明細書

データベース構築装置、データベース検索装置、データベース装置、データベース構築方法、及びデータベース検索方法

技術分野

[0001] 本発明は、XMLなどの論理構造を有する構造化文書を管理するデータベース装置に関し、特に、大量の構造化文書を蓄積管理するデータベース構築装置とそれに蓄積された構造化文書を効率良く検索するデータベース検索装置に関する。

背景技術

[0002] 特開2002－202973号公報は、構造化文書を論理構造に基づいて登録し、論理構造を指定して全文検索する構造化文書管理装置を開示している。

[0003] 図33は、従来の構造化文書管理装置の構成図である。構造化文書入力部2402は登録対象の構造化文書を入力する。構造解析部2407は入力された構造化文書を木構造に解析する。検索エンジン2405内で、構造情報作成部2408は、各要素のタグ名（要素名）に名称IDを割り振り、データ格納部2406内の名称IDテーブル格納部2418に格納する。また、各要素のパス名称、すなわち、最上位階層から順にタグ名を連ねて記述した文字列には、パス名称IDを割り振り、パス名称インデックス格納部2416に格納する。各要素のパス階層、すなわち、パス名称の各階層の出現順序を連ねて記述した文字列には、パス階層IDを割り当て、パス階層インデックス格納部2417に格納する。なお、パス名称の各階層の出現順序は、同じ親要素を持つ同じタグ名の要素の中で何番目に出現した要素かを示す。実体（テキスト）を持つ要素（以下、「要素実体」と記載する）の場合は、各要素実体に対し、検索単位を一意に表す符合（以下、「検索単位識別子」と記載する）を割り当て、要素管理テーブル格納部2415に格納する。図34は、従来の構造化文書管理装置における要素管理テーブルの例を示す図である。図34においては、要素管理テーブル2501は、検索単位識別子2502をキーとして、文書番号2503、パス名称ID2504、パス階層ID2505、名称ID2506の組とで構成されている。

[0004] 次に、文字列索引作成部2409は、各要素実体の内容である文字列に対して、予

め定められた文字数の文字連鎖を取り出す。そして、文字列索引作成部2409は、この文字連鎖について、該当する検索単位識別子と、該文字連鎖先頭文字がその要素内容において何番目の文字かを表す番号（以下、「文字位置番号」と記載する）とを文字列索引格納部2419に格納する。図35Aは、構造化文書の例を示す。また、図35Bは、従来の構造化文書管理装置における文字列索引の例を示す図である。図35Bにおいて、文字列索引2602のレコード2606は、「検索単位識別子2604が“1”の要素の文字列中に、“構造”という文字連鎖2603が、文字位置番号2605が“1”、すなわち、要素の先頭から“1”文字目に存在する」ことを表す。

[0005] 次に、このようにして格納されたデータを用いた検索の概要を説明する。図36A－36Cを用いて、従来の構造化文書管理装置における検索処理の動作を説明する。図36Aは、検索条件の設定例を示す図である。図36Aにおいて、構造を指定した検索条件2701は、「パス名称が“／論文／書誌／タイトル”である要素に“構造化”という文字列が含まれる文書」、を示す。検索条件解析部2410は、パス名称インデックス格納部2416を参照して、検索条件のパス名称をパス名称ID“N2”に変換する（2702）。次に文字列索引検索部2411は、“構造化”から2文字連鎖“構造”と“造化”を取り出す。そして、文字列索引を参照して、“構造”と“造化”が連続して出現し、かつ、検索単位識別子が同一のエントリの検索単位識別子を求める（2703）。ここでは、文字列索引検索結果群として、図36Cに示すように、検索単位識別子“1”と“8”が求められたとして説明する。

[0006] 次に、構造照合部2412は、検索条件2702、2703の構造指定を満たす検索結果を求める。ここで、構造照合部2412は、文字列索引検索結果群として得られた検索単位識別子をキーにして、図36Bに示す要素管理テーブル2501を検索する。そして、パス名称IDが“N2”に一致するエントリを検索結果として決定する。検索結果を図36Cに示す。もし、検索条件がタグ名を指定した条件の場合には、構造照合部2412は、要素管理テーブルの名称IDが指定タグ名の名称IDと一致するエントリを検索結果とする。また、検索条件が、パス名称とパス階層をともに指定した条件の場合には、構造照合部2412は、要素管理テーブルのパス名称IDが指定したパス名称のパス名称IDと一致し、かつ、パス階層IDが指定したパス階層のパス階層IDと一致する

エントリを検索結果とする。

[0007] また、特開2004－310607号公報は、構造化文書に含まれる要素を階層構造上の位置と結び付けるインデクスを生成する文書管理装置を開示している。この文書管理装置は、階層構造上の位置までの探索経路が同じである要素、すなわち1の親ノードに対して複数の子ノードが存在するような構成の要素であっても、複数の要素それぞれを識別して管理することができる。

[0008] 上記従来の構造化文書管理装置は、まず文字列索引を参照して指定された文字列の出現する検索単位識別子を求めた後、検索単位識別子が指定された構造条件を満たすかどうかを、要素管理テーブルを参照して判定する。そのため、文書検索をするときに文字列検索条件を指定する必要があり、構造条件だけを指定した検索ができない。すなわち、構造条件だけを指定して検索するためには、全ての検索単位識別子について構造条件を満たすかどうかについて、要素管理テーブル全体をサーチして判定する。そのため、効率が非常に悪いという課題がある。

[0009] また、構造化文書データを蓄積する際に、全文検索のための検索インデクスデータに論理構造データを付加するデータ構造としている。そのため、構造条件だけを指定した検索に対して効率的な検索を可能とする構造の検索用データを構築することができない。

[0010] また、文字列索引は要素実体の内容文字列に対してのみ作成されるため、要素の属性値に対しては文字列検索することができない。

発明の開示

[0011] 本発明のデータベース構築装置は、構造化文書にユニークな文書番号を割り当てるとともに構造を解析する入力文書解析部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名IDを割り当てて要素名辞書に登録する要素名登録部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書に登録する祖先パス名登録部と、入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして要素出現情報格納部に登録し

、かつ、文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして祖先パス出現情報格納部に登録する出現情報登録部とを備える。

[0012] このデータベース構築装置では、構造化文書を登録蓄積する際に、要素の出現情報に基づいて適切な出現情報インデクスを生成する。したがって、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造条件だけを指定した様々な検索条件に対しても、本発明のデータベース構築装置は、所望の文書を効率良く検索することが可能な構造の検索用データを構築できる。

図面の簡単な説明

[0013] [図1]図1は本発明の実施の形態1におけるデータベース装置の構成を示すブロック図である。

[図2]図2は本発明の実施の形態1における文書登録処理の手順を示す流れ図である。

[図3]図3は本発明の実施の形態1における登録検索対象となる構造化文書の例を示す図である。

[図4]図4は本発明の実施の形態1における構造化文書の論理構造を解析した結果の例を示す図である。

[図5]図5は本発明の実施の形態1における祖先パス名を説明する図である。

[図6]図6は本発明の実施の形態1における要素名辞書の内容の例を示す図である。

[図7]図7は本発明の実施の形態1における祖先パス名辞書の内容の例を示す図である。

[図8]図8は本発明の実施の形態1における属性名辞書の内容の例を示す図である。

[図9]図9は本発明の実施の形態1における文字位置を説明する図である。

[図10A]図10Aは本発明の実施の形態1における要素出現情報を説明する図である。

[図10B]図10Bは本発明の実施の形態1における要素出現情報を説明する図である。

[図11]図11は本発明の実施の形態1における祖先パス出現情報を説明する図であ

る。

[図12A]図12Aは本発明の実施の形態1における属性出現情報を説明する図である。

[図12B]図12Bは本発明の実施の形態1における属性出現情報を説明する図である。

[図13]図13は本発明の実施の形態1におけるテキスト出現情報を説明する図である。

[図14]図14は本発明の実施の形態1における検索式の例を示す図である。

[図15]図15は本発明の実施の形態1におけるデータベース装置の検索処理の手順を示す流れ図である。

[図16A]図16Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図16B]図16Bは本発明の実施の形態1におけるデータベース装置の検索動作を説明する図である。

[図16C]図16Cは本発明の実施の形態1における検索結果を説明する図である。

[図17A]図17Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図17B]図17Bは本発明の実施の形態1におけるデータベース装置の検索動作を説明する図である。

[図17C]図17Cは本発明の実施の形態1における検索結果を説明する図である。

[図18A]図18Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図18B]図18Bは本発明の実施の形態1におけるデータベース装置の検索動作を説明する図である。

[図18C]図18Cは本発明の実施の形態1における検索結果を説明する図である。

[図19A]図19Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図19B]図19Bは本発明の実施の形態1におけるデータベース装置の検索動作を説

明する図である。

[図19C]図19Cは本発明の実施の形態1における検索結果を説明する図である。

[図20A]図20Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図20B]図20Bは本発明の実施の形態1におけるデータベース装置の検索動作を説明する図である。

[図20C]図20Cは本発明の実施の形態1における検索結果を説明する図である。

[図21A]図21Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図21B]図21Bは本発明の実施の形態1におけるデータベース装置の検索動作を説明する図である。

[図21C]図21Cは本発明の実施の形態1における検索結果を説明する図である。

[図22A]図22Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図22B]図22Bは本発明の実施の形態1におけるデータベース装置の検索動作を説明する図である。

[図22C]図22Cは本発明の実施の形態1における検索結果を説明する図である。

[図23A]図23Aは本発明の実施の形態1における検索条件の例を説明する図である。

[図23B]図23Bは本発明の実施の形態1におけるデータベース装置の検索動作を説明する図である。

[図23C]図23Cは本発明の実施の形態1における検索結果を説明する図である。

[図24]図24は本発明の実施の形態2における空要素順の説明に用いる図である。

[図25A]図25Aは本発明の実施の形態2における部分祖先パス名を説明する図である。

[図25B]図25Bは本発明の実施の形態2における祖先パス名辞書の内容を示す図である。

[図25C]図25Cは本発明の実施の形態2における祖先パス名ID列を説明する図であ

る。

[図26]図26は本発明の実施の形態2における要素出現情報を説明する図である。

[図27]図27は本発明の実施の形態2における祖先パス出現情報を説明する図である。

[図28]図28は本発明の実施の形態2における検索式の例を示す図である。

[図29A]図29Aは本発明の実施の形態2における検索動作を説明する図である。

[図29B]図29Bは本発明の実施の形態2における検索結果を説明する図である。

[図30]図30は本発明の実施の形態3におけるデータベース装置の構成を示すブロック図である。

[図31]図31は本発明の実施の形態3におけるデータベース装置の文書登録処理の手順を示す流れ図である。

[図32]図32は本発明の実施の形態3におけるグループ化された要素出現情報を説明する図である。

[図33]図33は従来の構造化文書管理装置の構成図である。

[図34]図34は従来の構造化文書管理装置における要素管理テーブルの例を示す図である。

[図35A]図35Aは従来の構造化文書管理装置で処理する構造化文書の例を示す図である。

[図35B]図35Bは従来の構造化文書管理装置における文字列索引の例を示す図である。

[図36A]図36Aは従来の構造化文書管理装置における検索条件の例を説明する図である。

[図36B]図36Bは従来の構造化文書管理装置における検索動作を説明する図である。

[図36C]図36Cは従来の構造化文書管理装置における検索結果を説明する図である。

符号の説明

[0014] 101 構造化文書群

102　入力文書解析部

103　要素名登録部

104　祖先パス名登録部

105　属性名登録部

106　出現情報登録部

107　要素名辞書

108　祖先パス名辞書

109　属性名辞書

110　出現位置索引

111　要素出現情報格納部

112　祖先パス出現情報格納部

113　属性出現情報格納部

114　テキスト出現情報格納部

115　検索式

116　検索条件入力部

117　検索条件解析部

118　出現情報取得部

119　検索結果出力部

120　検索結果

2101, 2102, 2103, 2104, 2105, 2106, 2107, 3201　検索式

3401　出現情報グループ化部

発明を実施するための最良の形態

[0015]　(実施の形態1)

図1は、本発明の実施の形態1におけるデータベース装置の構成を示すブロック図である。図1において、本実施の形態におけるデータベース装置は、データベースに登録する構造化文書群101を入力し、入力された構造化文書群101の各文書についてユニークな文書番号を割り振るとともに論理構造を解析する入力文書解析部102、入力文書解析部102の解析結果から、文書に出現する要素名に対してユニーク

な識別子（以下、「要素名ID」と記載する）を割り当てて要素名辞書107に登録する要素名登録部103、入力文書解析部102の解析結果から、文書に出現する祖先パス名（着目要素の祖先要素の要素名を最上位階層から順にスラッシュで区切って並べた文字列で、着目要素自身の要素名は含まない）に対してユニークな識別子（以下、「祖先パス名ID」と記載する）を割り当てて祖先パス名辞書108に登録する祖先パス名登録部104、入力文書解析部102の解析結果から、文書に出現する属性名に対してユニークな識別子（以下、「属性名ID」と記載する）を割り当てて属性名辞書109に登録する属性名登録部105、入力文書解析部102の解析結果から、出現位置索引110の要素出現情報格納部111、祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114に4種の出現情報を登録する出現情報登録部106を備える。さらにデータベース装置は、上述した要素名IDとそれに対応する要素名が記録された要素名辞書107、祖先パス名IDとそれに対応する祖先パス名が記録された祖先パス名辞書108、属性名IDとそれに対応する属性名が記録された属性名辞書109、4種の出現情報がそれぞれ格納されている出現位置索引110を備える。この出現位置索引110は、要素出現情報格納部111、祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114を備える。要素出現情報格納部111は、各要素の出現する文書番号、文字位置、文字数、祖先パス名ID、分岐順の情報を、要素名IDをキーにして格納し、祖先パス出現情報格納部112は、各要素の出現する文書番号、文字位置、文字数、要素名ID、分岐順の情報を、その要素の祖先パス名IDをキーにして格納し、属性出現情報格納部113は、各属性の出現する文書番号、文字位置、文字数、要素名ID、祖先パス名ID、分岐順の情報を、属性名IDをキーにして格納し、テキスト出現情報格納部114は、要素内のテキストから切り出した部分文字列、および要素の持つ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名ID、要素名ID、属性名ID、分岐順の情報を、部分文字列をキーにして格納する。さらに、データベース装置は、検索式115を受け付ける検索条件入力部116、検索条件入力部116に与えられた検索式を解析し、内部条件に変換して出現情報取得部118に出力する検索条件解析部117、検索条件解析部117の出力した内部条件に応じて、出現

位置索引110に格納された4種の出現情報から適切な情報を選択して取得し、検索条件にマッチする結果データ集合を求める出現情報取得部118、結果データ集合を適切な形式で検索結果120として出力する検索結果出力部119を備える。

[0016] 本実施の形態におけるデータベース装置の動作について説明する。

[0017] はじめに、文書登録するデータベース構築処理について説明する。図2は、本発明の実施の形態1における文書登録処理の手順を示す流れ図である。

[0018] ステップ2201において、入力文書解析部102は、構造化文書群101から構造化文書を1つ読み込んで、文書毎にユニークな文書番号を割り振る。

[0019] ステップ2202において、入力文書解析部102は、この文書の論理構造を解析する。図3は、本発明の実施の形態1における登録検索対象となる構造化文書の例を示す図である。図3に示した構造化文書101aは、最上位階層にブック(book)要素を持ち、さらに、ブック要素はタイトル(title)要素と2つのチャプタ(chapter)要素を含む。タイトル要素は、要素実体の文字列"文書検索"を含み、さらに、1つ目のチャプタ要素は別のタイトル要素と2つのセクション(section)要素および属性値が"歴史"であるキーワード(keyword)属性を持つ。また、入力文書解析部102が構造化文書101aを木構造に解析した結果を図4に示す。図4は、本発明の実施の形態1における構造化文書の論理構造を解析した結果を示す図である。図4において、木構造300の四角い枠は要素301～303を表し、枠内に記された文字列は要素名304を示す。また、楕円の点線枠は属性305を表し、枠内に記された文字列は属性名306(アップデート(update))を示す。

[0020] 木構造300の最上位階層の要素301から着目要素に至る経路の途中に存在する要素(以下、「祖先要素」と記載する)の要素名をスラッシュ記号"／"で区切り順に並べたものを「パス名」と呼ぶ。パス名の末尾部分、すなわち、着目要素自身の要素名を除いた部分を「祖先パス名」と呼ぶ。図5は、本発明の実施の形態1における祖先パス名を説明する図である。図5において、図4で網掛けを施した要素302のパス名701は、祖先パス名702、要素名703で構成される。

[0021] また、図4において、各要素の右肩に記した文字列を「分岐順」と呼ぶ。例えば、要素302の分岐順307は「1／2／3」である。分岐順は、パス名中の各要素について、

同じ親要素を持つ同じ要素名の要素の中で何番目に出現したかを示す番号を順に並べたものである。図4で網掛けした要素302とその左隣の要素303とは、パス名は同じであるが、分岐順307、308は異なる。なお、分岐順の表記方法はこれに限らない。例えば、1以外の値を持つ階層の深さとその値を並べる方法でもよい。この方法で分岐順307を表記すると「2:2, 3:3」となる。これは、深さ1の値は「1」なので省略し、深さ2の値が「2」であり、深さ3の値が「3」であるためである。同じ要素名の兄弟要素がめったに現れない文書、すなわち、分岐順の値がほとんど「1」であるような文書を格納する場合には、この表記方法の方が出現位置索引ファイルのサイズを小さくできる。

[0022] ステップ2203において、要素名登録部103は、着目要素の要素名が要素名辞書107に登録済みかどうかを調べる。もし、それが登録済みであれば対応する要素名IDを取得し、登録されていなければ新たに要素名ID(>0)を割り当てて、要素名と要素名IDを要素名辞書107に登録する。図6に、図3で示した構造化文書101aを登録処理した後における要素名辞書107の内容の例(407)を示す。

[0023] ステップ2204において、祖先パス名登録部104は、着目要素の祖先パス名が祖先パス名辞書108に登録済みかどうかを調べる。もし、それが登録済みであれば対応する祖先パス名IDを取得し、登録されていなければ新たに祖先パス名ID(>0)を割り当てて、祖先パス名を祖先パス名辞書108に登録する。図7に、図3で示した構造化文書101aを登録処理した後における祖先パス名辞書108の内容の例(408)を示す。

[0024] ステップ2205において、もし、着目要素が属性を持っていれば、ステップ2206へ進み、持っていなければ、ステップ2207へ進む。

[0025] ステップ2206において、属性名登録部105は、着目要素の各属性の属性名が属性名辞書109に登録済みかどうかを調べる。もし、それが登録済みであれば対応する属性名IDを取得し、登録されていなければ新たに属性名ID(>0)を割り当てて、属性名を属性名辞書109に登録する。ここで、図8に、図3で示した構造化文書101aを登録処理した後における属性名辞書109の内容の例(409)を示す。

[0026] ステップ2207において、出現情報登録部106は、着目要素に関する要素出現情

報を、要素名IDをキーとして要素出現情報格納部111に登録する。要素出現情報は、次の5種類の値の組、すなわち、文書番号、着目要素(子孫要素も含む)に含まれる(タグ以外の)テキストの先頭文字位置および文字数、祖先パス名ID、分岐順の値の組から構成される。図9は本実施の形態におけるデータベース装置での文字位置の数え方を説明する図である。図9において、テーブル410は、タグを除く当該文書内の全てのテキストをつなげた文字列において、各文字411の文字位置412を示す。なお、先頭の文字位置は「0」とする。図10A－10Bは、本発明の実施の形態1における要素出現情報を説明する図である。図10Bにおいて、図4で網掛けを施したセクション要素302の要素実体304は、先頭文字321の文字位置が「115」であり、要素実体322全体の文字数が「40」である。セクション要素302に関する要素出現情報501を図10Aに示す。図10Aにおいて、セクション要素302の要素名ID(502)は「4」、文書番号(503)は「1」である。そして、セクション要素302は、「115」文字目(文字位置504)から始まる長さ「40」文字(文字数505)の要素実体を含む。セクション要素302の祖先パス名ID(506)は「3」、分岐順(507)は「1／2／3」である。なお、祖先パス名ID506が「3」の祖先パス名は「／book／chapter」である。

[0027] ステップ2208において、出現情報登録部106は、着目要素に関する祖先パス出現情報を、祖先パス名IDをキーとして祖先パス出現情報格納部112に登録する。この祖先パス出現情報は、次の5種類の値の組、すなわち、文書番号、着目要素(子孫要素も含む)に含まれる(タグ以外の)テキストの先頭文字位置および文字数、要素名ID、分岐順の値の組で構成する。図11は、本発明の実施の形態1における祖先パス出現情報を説明する図である。図11において、図4の網掛けを施した要素302に関する祖先パス出現情報の内容511を示している。図10Aと図11に示すように、同一要素に関する要素出現情報と祖先パス出現情報は、キーとなる項目が要素名ID502であるか、あるいは、祖先パス名ID506であるか、という点が異なるのみである。

[0028] ステップ2209において、もし、着目要素が属性を持っているならば、ステップ2210に進み、着目要素が属性を持っていなければ、ステップ2211へ進む。

[0029] ステップ2210において、出現情報登録部106は着目要素の各属性に関する属性

出現情報を、属性名IDをキーとして属性出現情報格納部113に登録する。属性出現情報は、次の6種類の値の組、すなわち、文書番号、属性値の先頭文字位置および文字数、祖先パス名ID、要素名ID、分岐順の値の組から構成される。図12A－12Bは、本発明の実施の形態1における属性出現情報を説明する図である。図12Bにおいて、図4で網掛けを施したセクション要素302はアップデート属性305を含み、そのアップデート属性305の属性値350は、先頭文字351の文字位置351が「115」であり、属性値305全体の文字数352が「6」である。なお、属性出現情報における、属性値の先頭文字の文字位置は、図12Bに示すように、仮想的に着目要素322（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字321の文字位置と同じ値とする。セクション要素302のアップデート属性305に関する属性出現情報521を図12Aに示す。図12Aにおいて、属性名ID（522）は「2」、文書番号（503）は「1」である。そして、アップデート属性305は、「115」文字目（文字位置504）から始まる長さ「6」文字（文字数505）の属性値を持つ。アップデート属性305の所属する要素の祖先パス名ID（506）は「3」、要素名ID（502）は「4」、分岐順（507）は「1／2／3」である。なお、属性名IDが「2」の属性名は「update」であり、祖先パス名ID506が「3」の祖先パス名は「／book／section」である。また、要素名ID502が「4」の要素名は「section」である。

[0030] ステップ2211において、出現情報登録部106は、着目要素の実体内容のテキストから部分文字列を切り出す。そして、テキスト出現情報を、切り出した部分文字列をキーとしてテキスト出現情報格納部114に登録する。このとき、属性値と区別するため、属性名IDには常に0を格納する。テキスト出現情報は、次の6種類の値の組、すなわち、文書番号、切り出された部分文字列の先頭文字位置、祖先パス名ID、要素名ID、属性名ID、分岐順の値の組から構成される。

[0031] ステップ2212において、もし、着目要素が属性を持っているならば、ステップ2213に進み、属性を持っていなければ、ステップ2214に進む。

[0032] ステップ2213において、出現情報登録部106は、着目要素が持つ各属性の属性値文字列から部分文字列を切り出す。そして、テキスト出現情報格納部114に部分文字列をキーとして登録する。なお、属性値は図11に示す位置に仮想的に出現して

いるとして、属性出現情報と同様に、文字位置を算出する。また、ステップ2213では、ステップ2211での処理とは異なり、属性名IDには、着目している属性の属性名ID(＞0)を格納する。図13は、本発明の実施の形態1におけるテキスト出現情報を説明する図である。図13において、テキスト出現情報531(一部分)は、図4で網掛けを施したセクション要素302の要素実体(テキスト)と、セクション要素302のアップデート属性305の属性値についてのテキスト出現情報を含む。出現情報レコード1201は、セクション要素302の要素実体の例を示す。セクション要素302の要素実体の部分文字列(532)“極大”は、文書番号(503)が「1」の文書の「118」文字目(文字位置504)に現れる。そして、部分文字列の含まれる要素、すなわちセクション要素302の祖先パス名ID(506)は「3」、要素名ID(502)は「4」、分岐順(507)は「1／2／3」である。なお、祖先パス名ID506が3の祖先パス名は「／book／section」であり、要素名ID502が4の要素名は「chapter」である。ここで、部分文字列532が属性値であるか否かは、属性名ID522に応じて判別できる。ここでは、もし、属性名IDが「0」であれば、部分文字列532は属性値であると判別する。また、出現情報レコード1202は、セクション要素302におけるアップデート属性305の属性値の例を示す。アップデート属性305の属性値の部分文字列(532)“00”は、文書番号(503)が「1」の文書の「116」文字目(文字位置504)に現れる。そして、部分文字列の含まれる属性の要素、すなわちセクション要素302の祖先パス名IDは「3」、要素名ID(502)は「4」、分岐順(507)は「1／2／3」である。またその要素に属する属性名ID(522)は「2」である。なお、祖先パス名IDが「3」の祖先パス名は「／book／section」、要素名IDが「4」の要素名は「chapter」、属性名IDが「2」の属性名は「update」である。

[0033] ステップ2214において、この文書に出現する全ての要素について処理が終わったかどうかを調べ、もし未処理の要素が残っていればステップ2203に戻って処理を繰り返す。

[0034] ステップ2215において、全ての入力文書に対して処理が終わったかどうかを調べ、未処理の文書が残っていればステップ2201に戻って処理を繰り返す。

[0035] 以上のようにして、本実施の形態におけるデータベース装置は、文書登録し、データベース構築処理を完了する。

[0036] 次に、本実施の形態におけるデータベース装置が登録済みの文書群を検索する処理に関して説明する。

[0037] 図14は、本発明の実施の形態1における検索式の例を示す図である。これらの検索式2101－2107はW3C(World Wide Web Consortium)の勧告として公開されているエックスパス(XPath)言語で記述されている。なお、エックスパス言語の詳細な仕様はURL"http://www.w3.org/TR/xpath"に記載されている。

[0038] 検索式2101は「最上位階層のブック要素の子のチャプタ要素の子であるタイトル要素」を表す。検索式2102は「最上位階層のブック要素の子のチャプタ要素のいずれかの子要素」を表す。検索式2103は、「いずれかの階層にあるタイトル要素」を表す。検索式2104は「最上位階層のブック要素の子のチャプタ要素の子の2番目のセクション要素」を表す。検索式2105は、「最上位階層のブック要素の子のチャプタ要素の子のセクション要素のアップデート属性」を表す。検索式2106は、「最上位階層のブック要素の子のチャプタ要素の子のセクション要素で、かつ要素実体内容に"極大単語"という文字列を含む要素」を表す。検索式2107は、「最上位階層のブック要素の子のチャプタ要素の子のセクション要素のアップデート属性で、かつその属性値に"2004"という文字列を含む属性」を表す。

[0039] 次に、それぞれの検索式に対して、本実施の形態におけるデータベース装置が検索処理する動作を順に説明する。

[0040] (検索式2101の場合)

まず、検索式2101を検索条件として与えた場合の動作について説明する。図15は、本発明の実施の形態1におけるデータベース装置の検索処理の手順を示す流れ図である。

[0041] ステップ2301において、検索条件入力部116は検索式2101を入力する。

[0042] ステップ2302において、検索条件解析部117は、図16Aに示すように、入力された検索式2101を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID＝3かつ要素名ID＝2」に変換する。そして結果を出現情報取得部118に出力する。

[0043] ステップ2303において、出現情報取得部118は、出現位置索引110を参照し、要

素出現情報格納部111における要素名ID＝2のエントリ数Nを取得する。

[0044] ステップ2304において、出現情報取得部118は、出現位置索引110を参照し、祖先パス出現情報格納部112における祖先パス名ID＝3のエントリ数Mを取得する。

[0045] ステップ2305において、出現情報取得部118は、取得したエントリ数Nとエントリ数Mとを比較する。もし、N＜Mならば、ステップ2306へ進み、そうでなければ、ステップ2310へ進む。図16Bは、要素出現情報格納部111における要素名ID＝2のエントリ1301、図17Bは祖先パス出現情報格納部112における祖先パス名ID＝3のエントリ1401の例を示す。図16Aに示した例では、N＝8、M＝12である。この場合、N＜Mとなり、ステップ2306へ進む。図16Bの要素出現情報格納部111を選択する。

[0046] ステップ2306において、出現情報取得部118は、要素出現情報格納部111の要素名ID＝2のエントリ1301から1つ取得する。

[0047] ステップ2307において、出現情報取得部118は、このエントリの祖先パス名IDが3であるかどうかを調べる。そして、もし祖先パス名IDが3であればステップ2308へ進み、そうでなければ、ステップ2309へ進む。

[0048] ステップ2308において、出現情報取得部118は、このエントリのデータを結果データ集合1302に追加する。図16Cに結果データ集合を示す。結果データ集合1302の各データは、例えば、(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)のような形式で格納される。

[0049] ステップ2309において、出現情報取得部118は、Nエントリ全てについて処理したか調べる。もし、まだ処理していないエントリがあればステップ2306に戻って処理を繰り返す。

[0050] 次に、ステップ2305において、出現情報取得部118は、N＜Mでないと判定した場合には、ステップ2310へ進む。そして、出現情報取得部118は、図17Bに示すように、祖先パス出現情報格納部112における祖先パス名ID＝3の各エントリ1401を調べる。そして、出現情報取得部118は、要素名IDが2であるものを求め、図17Cに示すように結果データ集合1402に追加する(ステップ2310～ステップ2313)。

[0051] ステップ2314において、出現情報取得部118は、求められた結果データ集合を検索結果出力部119に出力する。検索結果出力部119は求めた結果データ集合の文

書実体を取得するなどして適切な形式で検索結果を出力する。

[0052] このように、本実施の形態におけるデータベース装置は、検索式2101に対しては、要素出現情報格納部111における指定した要素名IDのエントリから指定した祖先パス名IDを持つものを選ぶ処理と、祖先パス出現情報格納部112における指定した祖先パス名IDのエントリから指定した要素名IDを持つエントリを選ぶ処理のどちらか、エントリ数の少ない方を選択する。そのため、検索対象構造化文書群の論理構造の特性に応じて処理量を抑えることができ、所望の文書を効率良く検索できる。

[0053] (検索式2102の場合)

次に、検索条件入力部116に検索式2102を入力した場合の動作について説明する。検索条件解析部117は、図18Aに示すように、検索式2102を解析し、祖先パス名辞書108を参照して内部条件「祖先パス名ID＝3」に変換する。そして、結果を出現情報取得部118に出力する。出現情報取得部118は、出現位置索引110を参照し、図18Bに示すように祖先パス出現情報格納部112における祖先パス名ID＝3の全てのエントリ1501を求める。そして、例えば(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)のような形式で、図18Cに示すように、結果データ集合1502として検索結果出力部119に出力する。検索結果出力部119は求めた結果データ集合1502の文書実体を取得するなどして適切な形式で検索結果を出力する。

[0054] このように、本実施の形態におけるデータベース装置は、検索式2102に対しては、祖先パス出現情報格納部112における指定した祖先パス名IDのエントリを取得するだけで良いため、所望の文書を効率良く検索できる。

[0055] (検索式2103の場合)

次に、検索条件入力部116に検索式2103を入力した場合の動作について説明する。検索条件解析部117は、図19Aに示すように、検索式2103を解析し、要素名辞書107を参照して内部条件「要素名ID＝2」に変換する。そして、結果を出現情報取得部118に出力する。出現情報取得部118は、出現位置索引110を参照し、図19Bのように要素出現情報格納部111における要素名ID＝2の全てのエントリ1601を求める。そして、例えば(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)のような形式で、図19Cに示すように、結果データ集合1602を検索結果出力部119に

出力する。検索結果出力部119は求められた結果データ集合1602の文書実体を取得するなどして適切な形式で検索結果を出力する。

[0056] このように、本実施の形態におけるデータベース装置は、検索式2103に対しては、要素出現情報格納部111における指定した要素名IDのエントリを取得するだけで良いため、所望の文書を効率良く検索することができる。

[0057] （検索式2104の場合）

次に、検索条件入力部116に検索式2104を入力した場合の動作について説明する。検索条件解析部117は図20Aに示すように、検索式2104を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID＝3かつ要素名ID＝4かつ分岐順＝＊／＊／2」に変換する。そして、結果を出現情報取得部118に出力する。分岐順のアスタリスク「＊」の部分はどんな数字でもマッチすることを表す。出現情報取得部118は、出現位置索引110を参照して、要素出現情報格納部111における要素名ID＝4のエントリ数Nと祖先パス出現情報格納部112における祖先パス名ID＝3のエントリ数Mとを求める。そして、エントリ数NとMとを比較し、少ない方を選択する。もし、N＜Mでなければ、図20Bに示すように祖先パス出現情報格納部112における祖先パス名ID＝3の各エントリ1701を調べる。要素名IDが4であり、かつ分岐順が「＊／＊／2」であるエントリのデータを求める。そして、結果データ集合1702として、図20Cに示すように、例えば（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）のような形式で検索結果出力部119に出力する。もし、N＜Mならば、図示しない要素出現情報格納部111における要素名ID＝4の各エントリを調べる。そして、祖先パス名IDが3であり、かつ分岐順が「＊／＊／2」であるエントリのデータを求め、結果データ集合1702として検索結果出力部119に出力する。検索結果出力部119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

[0058] このように、本実施の形態におけるデータベース装置は、検索式2104に対しては、要素出現情報格納部111における指定した要素名IDのエントリから指定した祖先パス名IDと分岐順を持つものを選ぶ処理と、祖先パス出現情報格納部112における指定した祖先パス名IDのエントリから指定した要素名IDと分岐順を持つものを選ぶ

処理の、どちらか、エントリ数の少ない方を選択する。これにより、検索の処理量を減らすことが可能となり、所望の文書を効率良く検索することができる。

[0059] （検索式2105の場合）

次に、検索条件入力部116に検索式2105を入力した場合の動作について説明する。検索条件解析部117は、図21Aに示すように、検索式2105を解析し、要素名辞書107、祖先パス名辞書108、属性名辞書109を参照して内部条件「祖先パス名ID＝3かつ要素名ID＝4かつ属性名ID＝2」に変換する。そして、結果を出現情報取得部118に出力する。出現情報取得部118は、出現位置索引110を参照し、図21Bに示すように属性出現情報格納部113における属性名ID＝2の各エントリ1801を調べる。そして、祖先パス名IDが3であり、要素名IDが4であるエントリのデータを求める。そして、出現情報取得部118は、図21Cに示すように、例えば（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）のような形式で結果データ集合1802として検索結果出力部119に出力する。検索結果出力部119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

[0060] このように、本実施の形態におけるデータベース装置は、検索式2105に対しては、属性出現情報格納部113における指定した属性名IDのエントリから指定した祖先パス名IDと要素名IDを持つものを選び、所望の文書を検索することが可能となる。

[0061] （検索式2106の場合）

次に、検索条件入力部116に検索式2106を入力した場合の動作について説明する。検索条件解析部117は、図22Aに示すように、検索式2106を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件「祖先パス名ID＝3かつ要素名ID＝4かつ要素内に文字列“極大単語”を含む」に変換する。そして、結果を出現情報取得部118に出力する。出現情報取得部118は、出現位置索引110を参照し、図22Bに示すように、テキスト出現情報格納部114における“極大”のエントリ1901と、“単語”のエントリ1902とで連接演算する。その際、文書番号が同一であることと“単語”が“極大”の2文字後方に位置することだけでなく、祖先パス名IDが3、かつ要素名IDが4、かつ属性名IDが0、かつ分岐順が同一であるかチェックして、条件を満たすエントリを求める。そして、出現情報取得部118は、図22Cに示すように、例えば（文

書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）のような形式で結果データ集合1903として検索結果出力部119に出力する。検索結果出力部119は、求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

[0062] このように、本実施の形態におけるデータベース装置は、検索式2106に対しては、テキスト出現情報格納部114における部分文字列のエントリ同士で連接演算する際に、祖先パス名IDおよび要素名IDが指定した値であって、分岐順が同一であり、かつ属性名IDが0であるもの（1904、1905）を選び、所望の文書を検索することが可能となる。

[0063] （検索式2107の場合）

次に、検索条件入力部116に検索式2107を入力した場合の動作について説明する。検索条件解析部117は、図23Aに示すように、検索式2107を解析し、要素名辞書107、祖先パス名辞書108、属性名辞書109を参照して内部条件「祖先パス名ID＝3かつ要素名ID＝4かつ属性名ID＝2かつ属性値に文字列“2004”を含む」に変換する。そして、結果を出現情報取得部118に出力する。出現情報取得部118は、出現位置索引110を参照し、図23Bに示すように、テキスト出現情報格納部114における“20”のエントリ2001と“04”のエントリ2002との間で、連接演算する。その際、出現情報取得部118は、文書番号が同一であることと“20”が“04”の2文字後方に位置することだけでなく、祖先パス名IDが3、かつ要素名IDが4、かつ属性名IDが2、かつ分岐順が同一であるかをチェックし、条件を満たすエントリを求める。そして、出現情報取得部118は、図23Cに示すように、例えば（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）のような形式で結果データ集合2003として検索結果出力部119に出力する。検索結果出力部119は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

[0064] このように、本実施の形態におけるデータベース装置は、検索式2107に対しては、テキスト出現情報格納部114における部分文字列のエントリ同士で連接演算する際に、祖先パス名IDおよび要素名IDが指定した値であって、分岐順が同一であり、かつ属性名IDが指定した値（＞0）であるもの（2004、2005）を選び、所望の文書を

検索することが可能となる。

[0065] 以上説明したように、本実施の形態におけるデータベース装置は、要素の出現情報を、要素名IDをキーにして格納した要素出現情報格納部と、要素の出現情報を、その要素の祖先パス名IDをキーにして格納した祖先パス出現情報格納部と、属性の出現情報を、属性名IDをキーにして格納した属性出現情報格納部とを設ける。そのため、このデータベース装置は、構造条件だけを指定した検索式に対しても効率良く所望の文書を検索することができる。

[0066] また、本実施の形態におけるデータベース装置は、要素実体のテキスト文字列および要素の持つ属性の属性値から切り出した部分文字列の出現情報を格納したテキスト出現情報格納部をさらに設ける。そのため、このデータベース装置は、要素実体のテキストに対してだけでなく属性値に対しても文字列検索できる。

[0067] なお、本実施の形態におけるデータベース装置は、データベース構築処理において、要素実体や属性値から固定長の2文字連鎖で部分文字列を切り出す説明したが、他の切り出し方法、例えば特開平8－249354号公報「文書検索装置および単語索引作成方法および文書検索方法」に記載の方法等でも構わない。

[0068] また、本実施の形態におけるデータベース装置は、データベース検索処理において、検索条件式をエックスパス式で与えるとして説明したが、同様の意味を表す他のクエリ言語で与えるとしても本発明を適用することは可能である。

[0069] このようにすることによって、本実施の形態におけるデータベース装置では、構造化文書の登録の際に、構造化文書に含まれる文書構造を示す要素名と祖先パス名と属性名の一覧と、それらの構造化文書中での出現位置情報のインデクスを生成する。そのため、このデータベース装置は、文字列検索条件と構造条件をともに指定した検索条件のみならず、構造だけを指定した様々な検索条件に対しても、所望の論理構造を持つ文書を効率良く検索するデータベースを構築できる。

[0070] また、要素実体のテキスト文字列に対してだけでなく、属性値に対しても文字列検索できる。

[0071] なお、本実施の形態におけるデータベース装置では、構造化文書を登録する際に、文書構造を解析して辞書データおよび出現位置索引データを構築して構造化文

書を登録する構成と、受け付けた文書構造を示す検索式に示される文書を辞書データおよび出現位置索引データに基づいて登録文書を効率的に検索する構成とを、同時に実現する形態とした。しかし、登録する機能のみの構成をデータベース構築装置として、あるいは検索のみの構成をデータベース検索装置として実現してもよい。

[0072] なお、本実施の形態におけるデータベース装置では、構造化文書を登録する際に、要素と祖先パスに対する辞書データならびに出現位置索引データを生成して登録する構成と、この構成に属性に対する辞書データならびに出現位置索引データを生成して登録する構成と、さらにこの構成に要素や属性値のテキストに対する出現位置索引データを生成して登録する構成とを同時に実現する形態とした。しかし、要素と祖先パスのみを対象として登録する構成、あるいは、この構成に属性を対象に加えて登録する構成、あるいは、さらにこの構成にテキストを対象に加えて登録する構成として実現してもよい。

[0073] (実施の形態2)

次に、本実施の形態2におけるデータベース装置の構成と動作について説明する。本実施の形態におけるデータベース装置は、図1に示した実施の形態1とほぼ同じ構成をしている。しかし、このデータベース装置は、次の点が実施の形態1とは異なっている。このデータベース装置は、祖先パス名登録部104が、文書に出現する各祖先パス名に対してではなく、祖先パス名をいくつかに分割した各部分祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書108に登録する。また、このデータベース装置は、出現情報登録部106が、各要素の出現する文書番号、文字位置、文字数、祖先パス名ID列、分岐順、空要素順の情報を、要素名IDをキーにして要素出現情報格納部111へ格納する。また、このデータベース装置は、各要素の出現する文書番号、文字位置、文字数、要素名ID、分岐順、空要素順の情報を、祖先パス名ID列をキーにして祖先パス出現情報格納部112へ格納する。また、このデータベース装置は、各属性の出現する文書番号、文字位置、文字数、要素名ID、祖先パス名ID列、分岐順、空要素順の情報を、属性名IDをキーにして属性出現情報格納部113へ格納する。また、このデータベース装置は、要素内のテキストから切

り出した部分文字列、および要素の持つ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名ID列、要素名ID、属性名ID、分岐順、空要素順の情報を、部分文字列をキーにしてテキスト出現情報格納部114へ格納する。

[0074] 本実施の形態におけるデータベース装置が文書登録しデータベース構築する処理の動作について図2を用いて説明する。なお、実施の形態1と同様の処理については説明を省略する。

[0075] ステップ2201において、入力文書解析部102は構造化文書を1つ読み込みユニークな文書番号を割り振る。

[0076] ステップ2202において、この構造化文書の論理構造を解析する。その際、実施の形態1の場合での処理に加え、各要素に関する「空要素順」の情報についても求める。ここで、「空要素」とは、子孫要素を含めて要素実体のテキストを全く持たない要素のことであり、「空要素順」とは、同じ親要素を持つ兄弟要素のうちで、先頭の要素であるかもしくは直前の兄弟要素が空要素でない要素の場合には1、それ以外の場合、すなわち、直前の兄弟要素が空要素である場合には、その空要素順の値に1を加えた値を、最上位階層から当該要素に至るまでの各階層において求めて並べたものである。

[0077] 図24は、本発明の実施の形態2における空要素順の説明する図である。図24において、文書の木構造310と空要素順の一例を示す。斜線模様の四角い枠は要素実体のテキストを含む要素2801、2804、2805を、無地の四角い枠は要素実体を含まない空要素2802、2803を、各要素の右肩に「1／2／3」の形式で記された文字列は、各要素の空要素順2806の情報を表す。

[0078] 兄弟要素2801～2804の空要素順が示す最初の2つの数字「1／2」は祖先要素の空要素順にあたる。兄弟要素に共通であり、末尾の数字nが兄弟要素毎に変わりうる。要素2801は兄弟要素の中の先頭要素であるのでn＝1となる。要素2802は直前の要素2801が空要素ではないのでn＝1となる。要素2803は直前の要素2802が空要素なので1を加えてn＝2となる。要素2804は直前の要素2803が空要素なのでさらに1を加えてn＝3となる。したがって、兄弟要素2801～2804の空要素順は

それぞれ、「1／2／1」、「1／2／1」、「1／2／2」、「1／2／3」となる。

[0079] なお、空要素順の表記方法はこれに限らない。例えば、1以外の値を持つ階層の深さとその値を並べて表記する方法でもよい。この方法で空要素順2806(「1／2／3」)を表記すると、「2:2, 3:3」となる。ここで、深さ1の値は「1」なので省略し、深さ2の値が「2」であり、深さ3の値が「3」である。そのため、空要素がほとんど現れない文書、すなわち、空要素順の値がほとんど「1」である文書を扱う場合には、後者の表記方法の方が出現位置索引ファイルのサイズを小さくできる。

[0080] ステップ2203において、実施の形態1と同様に、要素名登録部103は、着目要素の要素名について、要素名辞書107への登録処理をする。

[0081] ステップ2204において、祖先パス名登録部104は、着目要素の祖先パス名を3階層毎に分割し、分割後の各部分祖先パス名が祖先パス名辞書108に登録済みかどうかを調べる。もし、それが登録済みであれば対応する祖先パス名IDを取得し、それが登録されていなければ新たに祖先パス名ID(＞0)を割り当てて、祖先パス名辞書108に登録する。なお、祖先パス名の深さが3階層以下ならば、祖先パス名ID列は、実施の形態1の場合と同じように単一の祖先パス名IDとなる。

[0082] 図25Aは、本発明の実施の形態2における部分祖先パス名を説明する図、図25Bは、祖先パス名辞書の内容を示す図、図25Cは、祖先パス名ID列を説明する図である。図25Aにおいて、パス名2900より要素名2911を除いた祖先パス名2901「／A／B／C／A／B／C／A／B／C」は、さらに部分パス名「／A／B／C」(2913、2914)と「／A／B／」(2915)とに分解できる。ここで、図25Bに示すように、祖先パス名辞書108の内容2903に、祖先パス名2905「／A／B／C」、「／A／B」の祖先パスID2904が、それぞれ「83」、「25」と登録されている。この場合、図25Cに示すように、祖先パス名2901は、分解した各祖先パス名2905を示す祖先パスID2904と、記号「:」を用いて、祖先パス名ID列2902「83:83:25」のように表現できる。

[0083] このように、祖先パス名2901を分割して各部分祖先パス名2905に祖先パス名ID2904を割り当て、当該要素の祖先要素や他の要素との間で、登録済みの祖先パス名ID2904を共通に用いることができる。また、祖先パス名IDの重なる数を小さくでき、祖先パス名辞書108のサイズを小さくできる。

[0084] なお、本実施例では祖先パス名を3階層毎に分割する例を示したが、分割の方法はこれに限らない。例えば4階層毎に分割し、階層の深さによって分割幅を変化させるようにしても構わない。また、祖先パス名ID列の区切り文字として記号":"を用いたが、他の区切り文字でも構わない。

[0085] もし、着目要素が属性を持っているならば、ステップ2205～ステップ2206において、属性名登録部105は、実施の形態1と同様に、着目要素の各属性の属性名辞書109への登録処理をする。

[0086] ステップ2207において、出現情報登録部106は、着目要素に関する要素出現情報を、要素名IDをキーとして要素出現情報格納部111に登録する。要素出現情報は、次の6種類の値の組、すなわち、文書番号、着目要素(子孫要素も含む)に含まれる(タグ以外の)テキストの先頭文字位置および文字数、祖先パス名ID列、分岐順、空要素順の値の組から構成される。なお、「文字位置」は、タグを除く当該文書内の全てのテキストをつなげた文字列において先頭から何文字目にあたるかで表す。また、着目要素が空要素である場合には、着目要素以降に初めて現れる(タグ以外の)テキストの先頭文字位置を着目要素の先頭文字位置とみなす。要素出現情報の一例を図26に示す。図26は、本発明の実施の形態2における要素出現情報を説明する図である。実施の形態1と異なるのは、要素出現情報541の祖先パス名506に、単一の祖先パス名IDではなく1つ以上の祖先パス名IDを区切り文字で連ねた祖先パス名ID列が記録されることと、空要素順548の情報を含むことである。

[0087] ステップ2208において、出現情報登録部106は、着目要素に関する祖先パス出現情報を、祖先パス名ID列をキーとして祖先パス出現情報格納部112に登録する。祖先パス出現情報は、次の6種類の値の組、すなわち、文書番号、着目要素(子孫要素も含む)に含まれる(タグ以外の)テキストの先頭文字位置および文字数、要素名ID、分岐順、空要素順の値の組で構成する。祖先パス出現情報の一例を図27に示す。図27は、本発明の実施の形態2における祖先パス出現情報を説明する図である。実施の形態1と異なるのは、祖先パス出現情報551に空要素順548の情報を含むことと、祖先パス名ID506に、単一の祖先パス名IDではなく1つ以上の祖先パス名IDを区切り文字で連ねた祖先パス名ID列をキーとして登録することである。

[0088] もし、着目要素が属性を持っているならば、ステップ2209～ステップ2210において、出現情報登録部106は着目要素の各属性に関する属性出現情報を、属性名IDをキーとして属性出現情報格納部113に登録する。属性出現情報は、次の7種類の値の組、すなわち、文書番号、属性値の先頭文字位置および文字数、祖先パス名ID列、要素名ID、分岐順、空要素順の値の組から構成される。実施の形態1と異なるのは、属性出現情報の祖先パス名IDに単一の祖先パス名IDではなく1つ以上の祖先パス名IDを区切り文字で連ねた祖先パス名ID列を記録することと、空要素順の情報を含むことである。

[0089] ステップ2211において、出現情報登録部106は、着目要素の実体内容のテキストから部分文字列を切り出し、テキスト出現情報を、切り出した部分文字列をキーとしてテキスト出現情報格納部114に登録する。ただし、テキスト出現情報は属性値ではないので、属性名IDには常に値「0」を格納する。テキスト出現情報は、次の7種類の値の組、すなわち、文書番号、切り出した部分文字列の先頭文字位置、祖先パス名ID列、要素名ID、属性名ID、分岐順、空要素順の値の組から構成される。実施の形態1と異なるのは、テキスト出現情報の祖先パス名IDに単一の祖先パス名IDではなく、1つ以上の祖先パス名IDを区切り文字で連ねた祖先パス名ID列が記録されることと、空要素順の情報を含むことである。

[0090] もし、着目要素が属性を持っているならば、ステップ2212～ステップ2213において、出現情報登録部106は、着目要素が持つ各属性の属性値文字列から部分文字列を切り出し、テキスト出現情報格納部114に部分文字列をキーとして登録する。ステップ2211と同様に、実施の形態1と異なるのは、テキスト出現情報に単一の祖先パス名IDではなく1つ以上の祖先パス名IDを区切り文字で連ねた祖先パス名ID列を記録することと、空要素順の情報を含むことである。

[0091] 以降、ステップ2214～2215を実施の形態1と同様に実行し、文書登録してデータベース構築する。

[0092] 次に、登録済みの文書群を検索する処理に関して説明する。実施の形態1で示した検索式と同様の形式の検索式による索処理については、検索条件解析部117において、祖先パス名から祖先パス名IDを求めて内部条件に変換する処理を、祖先パ

ス名から祖先パス名ID列を求める処理に変更することで実現できる。すなわち、検索条件解析部117は祖先パス名を3階層毎に分割し、祖先パス名辞書108を参照して分割後の各部分祖先パス名に対応する祖先パス名IDを求め、それらの祖先パス名IDを順に区切り文字で区切って並べ祖先パス名ID列を求める。祖先パス名ID列の形式は、文書登録処理の説明で図25A－25Cに示した例と同様であり、祖先パス名の深さが3階層以下の場合には単一の祖先パス名IDとなる。実施の形態1では出現情報取得部118において祖先パス名IDで照合していた各処理を、祖先パス名ID列で照合するように変更することで、検索結果を求めることができる。

[0093] （検索式3201の場合）

図28は、本発明の実施の形態2における検索式の例を示す図である。図28に示す検索式3201は「最上位階層のA要素の子のB要素の子のX要素の兄弟要素であり、かつ、X要素より後ろに現れるY要素」を表す。検索条件入力部116より検索式3201を入力する。検索条件解析部117は、検索式3201を解析し、要素名辞書107、祖先パス名辞書108を参照して内部条件に変換し、出現情報取得部118に出力する。ただし、内部条件は、「C1かつ（C2またはC3）、ただし、Cx:{祖先パス名ID＝25かつ要素名ID＝10}、Cy:{祖先パス名ID＝25かつ要素名ID＝14}、C1:{CxとCyの文書番号が同一で、かつ分岐順が末尾以外等しい}、C2:{CxよりCyの方が文字位置の値が大きい}、C3:{CxとCyの文字位置の値が等しく、かつCxよりCyの方が空要素順の末尾の値が大きい}」である。ここで、祖先パス名「／A／B」に対応する祖先パス名IDは25であり、要素名「X」に対応する要素名IDは「10」であり、要素名「Y」に対応する要素名IDは「14」ある。ここで内部条件に条件C3を必要とする理由は、空要素とその直後に位置する要素では文字位置が同一になるため、前後関係を判断するために空要素順の値を比較しなければならないからである。

[0094] 本発明の実施の形態2における検索動作を説明する。出現情報取得部118は、出現位置索引110を参照し、図29Aに示すように、祖先パス出現情報格納部112における祖先パス名IDが25であるエントリのうち、要素名IDが10であるもの（Cx）、および要素名IDが14であるもの（Cy）を求める。続いて、C1かつ（C2またはC3）を満たすCx、Cyのエントリの組3301、3302を求める。そして、出現情報取得部118は、図

29Bに示すように、例えば、(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順, 空要素順)のような形式で結果データ集合3303として検索結果出力部119に出力する。検索結果出力部119は、求めた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

[0095] なお、CxおよびCyのエントリを求める際に、祖先パス出現情報格納部112における指定祖先パス名IDのエントリ数と、要素出現情報格納部111における指定要素名IDのエントリ数を比較して少ない方を選択して求めてもよい。

[0096] このようにして、本実施の形態におけるデータベース装置は、検索式3201に対して、祖先パス出現情報格納部112または要素出現情報格納部111を参照して求めた2つの要素の出現位置が同じ場合、すなわち、2つの要素が、空要素とその直後の要素の関係にある場合であっても、空要素順の情報を比較して、前後関係の曖昧さを排除し、正しく検索結果を求めることができる。

[0097] 以上説明したように、本実施の形態におけるデータベース装置は、祖先パス名登録部104が祖先パス名を分割し、分割後の各部分祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書108に登録する。そのため、祖先パス名辞書のサイズを小さくすることができる。

[0098] また、出現情報登録部106は要素出現情報格納部111と、祖先パス出現情報格納部112と、属性出現情報格納部113と、テキスト出現情報格納部114に空要素順の情報も格納する。そのため、本実施の形態におけるデータベース装置は、空要素とその直後の要素の開始文字位置が同じになるという前後関係の曖昧さを排除し、正しい検索結果を求めることができる。

[0099] このようにすることによって、本実施の形態におけるデータベース装置は、構造文書の要素にテキストが全く含まれない空要素である場合には、着目要素以降に初めて現れるテキストの先頭文字位置を着目要素の先頭文字位置とみなす。そのため、空要素の出現順を出現位置インデクスとして生成し、構造化文書に空要素が含まれる場合だけでなく空要素が連続して含まれる場合であっても、構造化文書構造の全文検索のみならず、空要素を含む文書構造を示す検索式に示される文書を効率的に検索することができる。

[0100] また、本実施の形態におけるデータベース装置は、祖先パス名を一定の条件で分割した部分パス名に基づいて祖先パス列として登録する。そのため、本実施の形態におけるデータベース装置は、部分パスを重複して蓄積することなく、結果的に祖先パス辞書のサイズを小さくでき、また、構造化対象を多く含む構造化文書であっても、文書構造を示す検索式に示される文書を効率的に検索できる。

[0101] なお、本実施の形態におけるデータベース装置は、構造化文書を登録する際に、文書構造を解析して辞書データおよび出現位置索引データを構築して構造化文書を登録する構成と、受け付けた文書構造を示す検索式に示される文書を辞書データおよび出現位置索引データに基づいて登録文書を効率的に検索する構成とを同時に実現する形態とした。しかし、構造化文書を登録する機能のみの構成、あるいは検索のみの構成として実現してもよい。

[0102] なお、本実施の形態におけるデータベース装置は、構造化文書を登録する際に、テキスト要素を持たない空要素に対応する出現位置索引データを生成して登録する構成と、祖先パス名をいくつかに分割した各部分祖先パス名に対する辞書データならびに出現位置索引データを生成して登録する構成とを同時に実現する形態とした。しかし、空要素のみを対象として登録する構成、あるいは、祖先パス名のみを対象として登録する構成として実現してもよい。

[0103] (実施の形態3)

次に、本実施の形態3におけるデータベース装置の構成と動作について説明する。図30は、本発明の実施の形態3におけるデータベース装置の構成を示すブロック図である。図30において、本実施の形態3におけるデータベース装置は、実施の形態2とほぼ同じ構成をしている。しかし、このデータベース装置は次の点が実施の形態2と異なっている。要素出現情報格納部111、祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114に格納されている情報をグループ化する出現情報グループ化部3401が追加されている。

[0104] 文書登録するデータベース構築処理の動作について説明する。図31は、本発明の実施の形態3におけるデータベース装置の文書登録処理の手順を示す流れ図である。図31において、ステップ2201～2215までの処理は実施の形態2の場合と同

じであるので、説明を省略する。

[0105] 最後のステップ3501において、出現情報グループ化部3401は、要素出現情報格納部111に同じ要素名IDをキーとして登録されているエントリ群の中で、文書番号と文字位置を除いた4種類の情報項目（文字数、祖先パス名ID、分岐順、空要素順）の値が全て共通しているエントリ同士を集め、それらのエントリの数が閾値（例えば、10エントリ）を超えていたらそれらのエントリをグループ化する。次に、出現情報グループ化部3401は、残りのエントリ群について、文書番号と文字位置を除いた4種類の情報項目（文字数、祖先パス名ID、分岐順、空要素順）のうち、いずれか3種類の情報項目の値が共通しているエントリ群を求め、そのエントリの数が閾値を超えていたらグループ化する。なお、複数のグループに属する可能性があるエントリは、エントリ数の最も多いグループに含める。さらに、出現情報グループ化部3401は、同様に、いずれか2種類の情報項目の値が共通するエントリのグループを作成する。さらに、出現情報グループ化部3401は、いずれか1種類の情報項目の値が共通するエントリのグループを作成し、最後に残ったエントリは共通情報項目無しのグループとして登録する。

[0106] 図32は、本発明の実施の形態3におけるグループ化された要素出現情報を説明する図である。図32において、要素名IDが14である要素出現情報がグループ化され、グループ情報と個々のエントリで構成されている。グループ情報3601～3604には、各グループに属するエントリ3605～3608に共通する情報項目の値と、個々のエントリへのリンク情報3615～3618を格納している。個々のエントリ3605～3608には、共通しない情報項目の値のみを格納している。

[0107] 第1のグループ情報3601は、当該グループに属する要素出現情報のエントリはいずれも（文字数＝10，祖先パス名ID＝100，分岐順＝“1／1／1”，空要素順＝“1／1／1”）という値を共通に持つ。当該グループに属する個々のエントリ3605は、それぞれの文書番号と文字位置だけを格納している。第2のグループ情報3602は、当該グループに属する要素出現情報のエントリはいずれも（祖先パス名ID＝200，分岐順＝“1／2／1”，空要素順＝“1／2／3”）という値を共通に持つが、記号「＊」で示される文字数の情報項目は共通な値ではないことを表す。個々のエントリ3606は

、文書番号、文字位置とともに文字数を格納する。第3のグループ情報3603は、当該グループに属する要素出現情報のエントリはいずれも(文字数＝8，祖先パス名ID＝150，空要素順＝“1／2”)という値を共通に持ち、記号「＊」で示される分岐順の情報項目は共通な値ではないことを表す。個々のエントリ3607は文書番号、文字位置とともに分岐順を格納する。第4のグループ情報3604は共通する情報項目がないグループであり、各エントリ3608に全ての情報項目を格納する。

[0108] 祖先パス出現情報格納部112、属性出現情報格納部113、テキスト出現情報格納部114に格納する各情報についても同様に、文書番号と文字位置以外の共通な値の情報項目を持つエントリ同士をグループ化し、文書登録するデータベース構築処理を完了する。

[0109] そのため、本実施の形態におけるデータベース装置の出現情報取得部118は、登録済みの文書群を検索する処理として、グループ化した各エントリの内容とグループ情報に基づいて全ての情報項目の値を復元し、実施の形態2と同様に検索結果を求める。

[0110] このようにして、本実施の形態におけるデータベース装置の出現情報グループ化部3401は、出現位置索引110に格納されるエントリ群をグループ化し、そのグループ内で共通する情報項目の値を括りだし、個々のエントリには格納しない。そのため、本実施の形態におけるデータベース装置は、索引サイズを減らすことができる。

[0111] このように、本実施の形態におけるデータベース装置は、各要素、祖先パスなどの出現位置情報について、ある条件下で情報項目の値が共通する部分をグループ化して、共通化できない部分とは異なる構造で格納する。そのため、共通する部分を重複して蓄積することなく、索引のサイズを小さくできる。

産業上の利用可能性

[0112] 本発明によるデータベース構築装置は、構造化文書を効率良く検索することが可能な構造の検索用データを構築でき、効率良く検索可能なデータベース装置等に有用である。

請求の範囲

[1] 構造化文書を管理するデータベース構築装置において、

構造化文書にユニークな文書番号を割り当てるとともに構造を解析する入力文書解析部と、

前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各要素名に対してユニークな要素名IDを割り当てて要素名辞書に登録する要素名登録部と、

前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書に登録する祖先パス名登録部と、

前記入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして要素出現情報格納部に登録し、かつ、着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして祖先パス出現情報格納部に登録する出現情報登録部と、

を有するデータベース構築装置。

[2] 前記入力文書解析部の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名IDを割り当てて属性名辞書に登録する属性名登録部を有し、

前記出現情報登録部が、前記入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして属性出現情報格納部に登録する請求項1に記載のデータベース構築装置。

[3] 前記出現情報登録部が、前記入力文書解析部の解析結果に基づいて、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名IDと要素名IDと属性名IDと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部に登録する

請求項1に記載のデータベース構築装置。

[4] 前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名IDと分岐順と空要素順の情報を少なくとも含む
請求項1に記載のデータベース構築装置。

[5] 前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名IDと分岐順と空要素順の情報を少なくとも含み、
前記属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順と空要素順の情報を少なくとも含む
請求項2に記載のデータベース構築装置。

[6] 前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名IDと分岐順と空要素順の情報を少なくとも含み、
前記テキスト出現情報は、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名IDと要素名IDと属性名IDと分岐順と空要素順の情報を少なくとも含む
請求項3に記載のデータベース構築装置。

[7] 前記祖先パス名登録部は、前記構造化文書に出現する各祖先パス名を1つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名IDを割り当てて前記祖先パス名辞書に登録する
請求項1に記載のデータベース構築装置。

[8] 前記要素出現情報格納部に同じ要素名IDをキーにして登録されている前記要素出現情報のエントリ群と、前記祖先パス出現情報格納部に同じ祖先パス名IDをキーにして登録されている前記祖先パス出現情報のエントリ群とに対して、文書番号と文字

位置以外の1つ以上の情報項目の値が共通するエントリ同士をグループ化する出現情報グループ化部を有する

請求項1に記載のデータベース構築装置。

[9] 構造化文書を管理するデータベース検索装置において、

構造化文書に出現する各要素名に対してユニークな要素名IDを登録した要素名辞書と、

前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを登録した祖先パス名辞書と、

前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして格納した要素出現情報格納部と、

前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして格納した、祖先パス出現情報格納部と、

検索式を入力するための検索条件入力部と、

前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換する検索条件解析部と、

前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報および、前記祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求める出現情報取得部と、

を有するデータベース検索装置。

[10] 属性名IDと対応する属性名の記録された属性名辞書と、

着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして格納した属性出現情報格納部とを有し、

前記検索条件解析部が、前記要素名辞書と前記祖先パス名辞書と前記属性名辞書とを参照して、前記検索条件入力部から入力された検索式を内部条件式に変換し、

前記出現情報取得部が、前記検索条件解析部の出力した内部条件式にしたがって

、前記要素出現情報格納部からの要素出現情報、前記祖先パス出現情報格納部からの祖先パス出現情報および、前記属性出現情報格納部からの属性出現情報から検索結果群を求める

請求項9に記載のデータベース検索装置。

[11] 要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名IDと要素名IDと属性名IDと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納部を有し、

前記出現情報取得部が、前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報、前記祖先パス出現情報格納部からの祖先パス出現情報、および、前記テキスト出現情報格納部からのテキスト出現情報から検索結果群を求める

請求項9に記載のデータベース検索装置。

[12] 前記出現情報取得部は、前記要素出現情報格納部における指定要素名IDのエントリ数と、前記祖先パス出現情報格納部における指定祖先パス名IDのエントリ数の大小を比較し、少ない方の出現情報を参照するようにして検索結果群を求める

請求項9乃至11のいずれかに記載のデータベース検索装置。

[13] 構造化文書を管理するデータベース構築方法において、

構造化文書にユニークな文書番号を割り当てるとともに構造を解析するステップと、

前記解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名IDを割り当てて要素名辞書に登録するステップと、

前記解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを割り当てて祖先パス名辞書に登録するステップと、

前記解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして要素出現情報格納部に、着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして祖先パス出現情報格納部にそれぞれ登録するステップと、

を有するデータベース構築方法。

[14] 前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名IDと分岐順と空要素順の情報を少なくとも含む、
請求項13に記載のデータベース構築方法。

[15] 前記祖先パス名辞書に登録するステップは、構造化文書に出現する各祖先パス名を1つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名IDを割り当てて登録するステップであり、
前記要素出現情報には、単一の祖先パス名IDの代わりに1つ以上の祖先パス名IDの列を含み、
前記祖先パス出現情報格納部には、単一の祖先パス名IDの代わりに1つ以上の祖先パス名IDの列をキーとして前記祖先パス出現情報を登録する、
請求項13記載のデータベース構築方法。

[16] 前記要素出現情報格納部に同一の要素名IDをキーとして登録され、文書番号と文字位置以外の情報項目の値が共通であるような前記要素出現情報のエントリ同士をグループ化し、前記祖先パス出現情報格納部に同一の祖先パス名IDをキーとして登録され、文書番号と文字位置以外の情報項目の値が共通であるような前記祖先パス出現情報のエントリ同士をグループ化するステップを有する
請求項13記載のデータベース構築方法。

[17] 構造化文書を管理するデータベース検索方法において、
構造化文書に出現する各要素名に対してユニークな要素名IDを登録した要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを登録した祖先パス名辞書と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして格納した要素出現情報格納部と、

前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして格納した、祖先パス出現情報格納部と、を備えたデータベース検索装置を用い、

検索式を入力するためのステップと、

前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換するステップと、

前記内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報および、前記祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求めるステップと、

を有するデータベース検索方法。

[18] 構造化文書を管理するデータベース装置において、

構造化文書に出現する各要素名に対してユニークな要素名IDを記憶する要素名辞書と、

前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを記憶する祖先パス名辞書と、

構造化文書にユニークな文書番号を割り当てるとともに構造を解析する入力文書解析部と、

前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各要素名に対してユニークな要素名IDを割り当てて前記要素名辞書に登録する要素名登録部と、

前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名IDを割り当てて前記祖先パス名辞書に登録する祖先パス名登録部と、

文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、要素名IDをキーとして記憶する要素出現情報格納部と、

文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名IDをキーとして記憶する祖先パス出現情報格納部と、

前記入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名IDと分岐順の情報を少なくとも含む要素出現情報を、前記着目要素の要素名IDをキーとして前記要素出現情報格納部に登録し、かつ、前記着目要素の出現する文書番号と文字位置と要素名IDと分岐順の情報を少なくとも含む祖先パス出現情報を、前記着目要素の祖先パス名IDをキーとして前記祖先パス出現情報格納部に登録する出現情報登録部とを具備するデータベース構築装置と、

検索式を入力する検索条件入力部と、

前記要素名辞書と前記祖先パス名辞書とを参照して、前記検索条件入力部で入力された検索式について要素名と祖先パス名とをそれぞれ要素名IDと祖先パス名IDとで表現した内部条件式に変換する検索条件解析部と、

前記要素出現情報格納部に記憶している要素出現情報、および、前記祖先パス出現情報格納部に記憶している祖先パス出現情報から、前記検索条件解析部で生成された前記内部条件式にあてはまる検索結果群データを抽出する出現情報取得部とを具備するデータベース検索装置と

を有するデータベース装置。

[19] 属性名IDと対応する属性名を記憶する属性名辞書と、

前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各属性名に対してユニークな属性名IDを割り当てて前記属性名辞書に登録する属性名登録部と、

文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして記憶する属性出現情報格納部とをさらに有し、

前記出現情報登録部は、さらに、前記入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名IDと要素名IDと分岐順の情報を少なくとも含む属性出現情報を、属性名IDをキーとして前記属性出現情報格納部に登録するようにし、

前記検索条件解析部は、さらに、前記属性名辞書を参照して、前記検索条件入力部で入力された検索式について、属性名を属性IDで表現した内部条件式に変換す

るようにし、

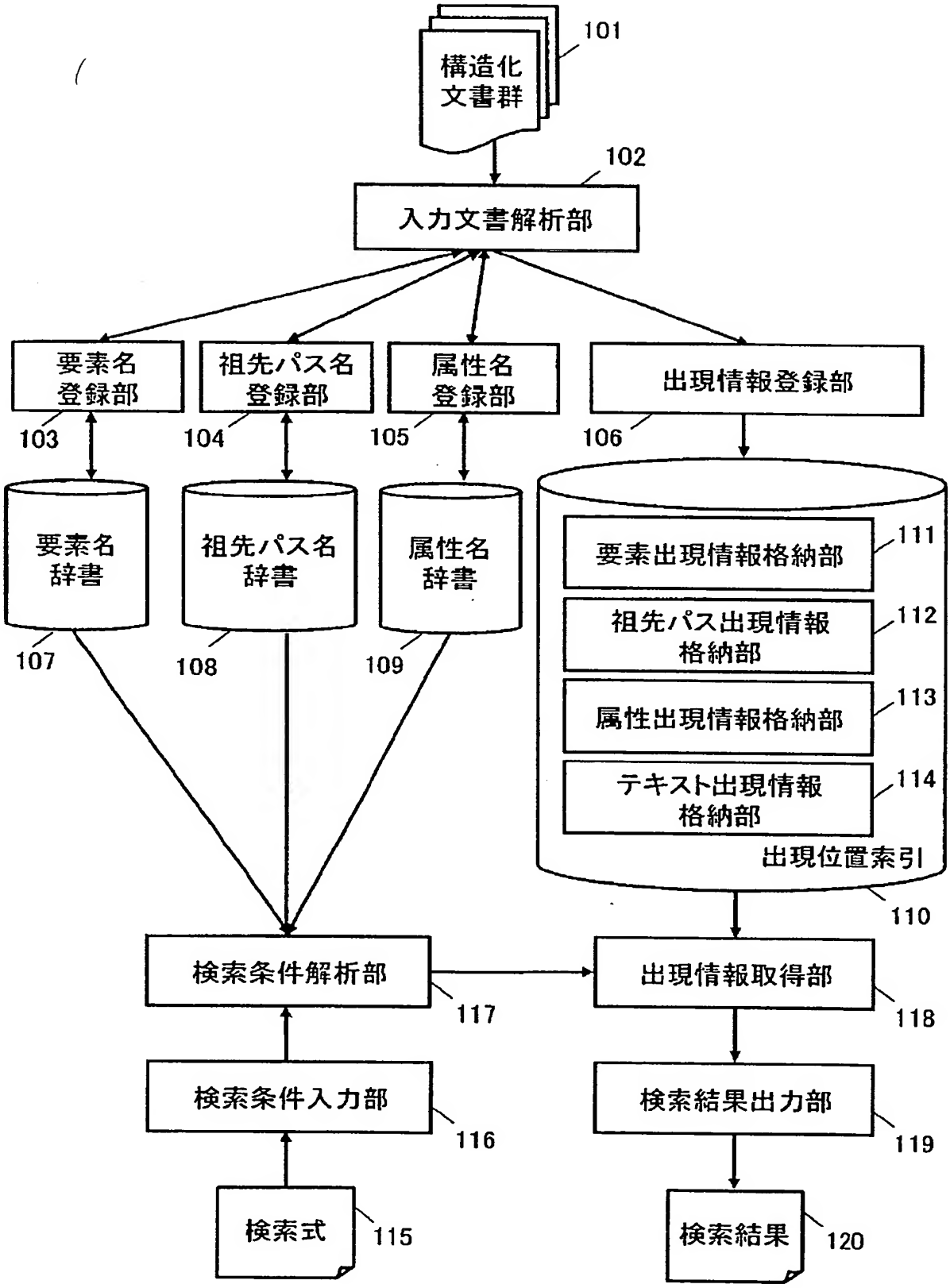
前記出現情報取得部は、さらに、前記要素出現情報格納部に記憶している要素出現情報と、前記祖先パス出現情報格納部に記憶している祖先パス出現情報と、前記属性出現情報格納部に記憶している属性出現情報とから前記検索条件解析部の出力した前記内部条件式にあてはまる検索結果群データを抽出する

請求項18に記載のデータベース装置。

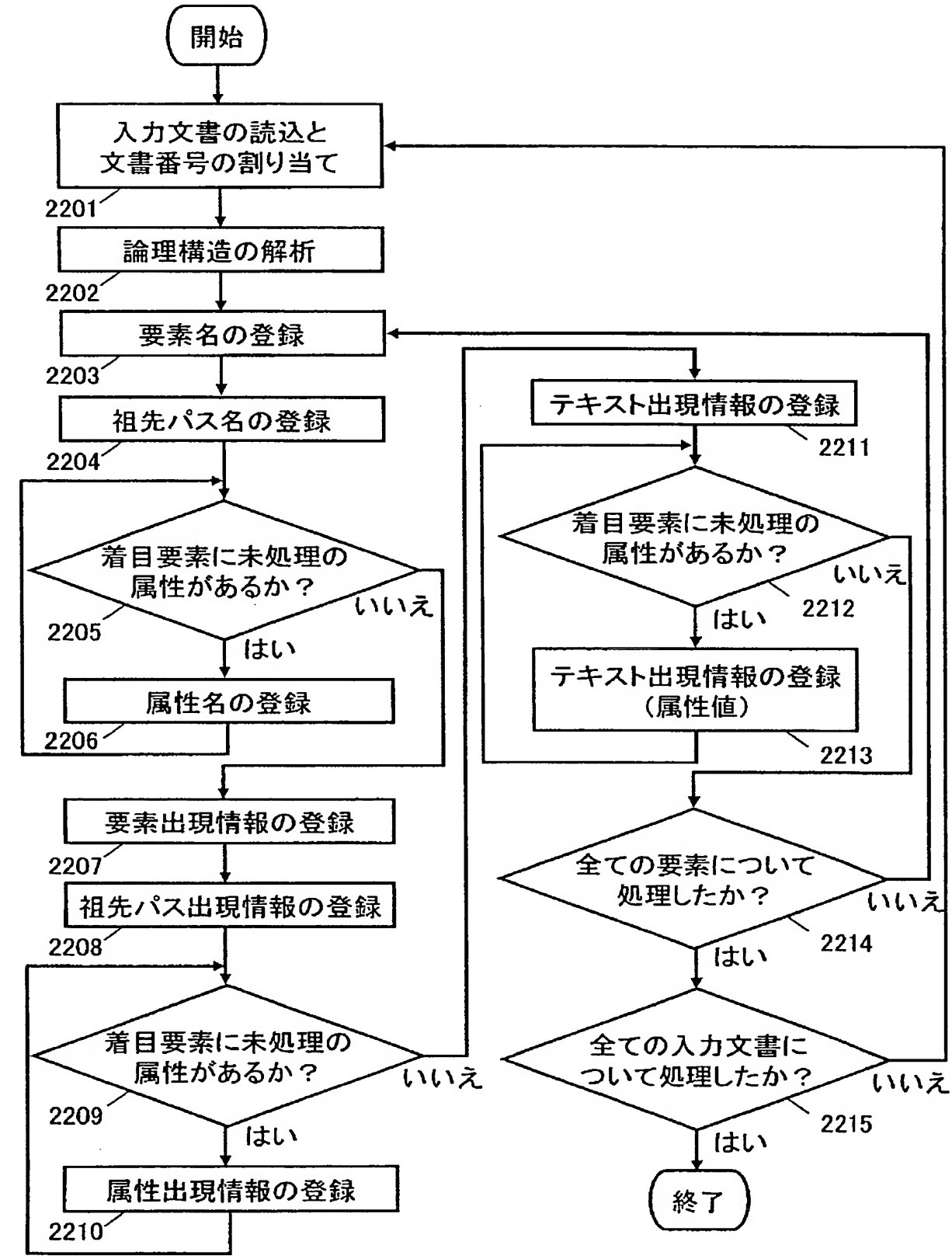
要 約 書

データベース装置は、要素の出現情報を、要素名IDをキーにして格納した要素出現情報格納部と、要素の出現情報を、その要素の祖先パス名IDをキーにして格納した祖先パス出現情報格納部と、属性の出現情報を、属性名IDをキーにして格納した属性出現情報格納部と、要素実体のテキスト文字列と要素の持つ属性の属性値に関する出現情報を、部分文字列をキーにして格納したテキスト出現情報格納部とを備える。このことによって、様々な検索条件での構造化文書を構造条件のみで検索でき、また、属性値に対する文字列検索できるデータベース装置を得る。

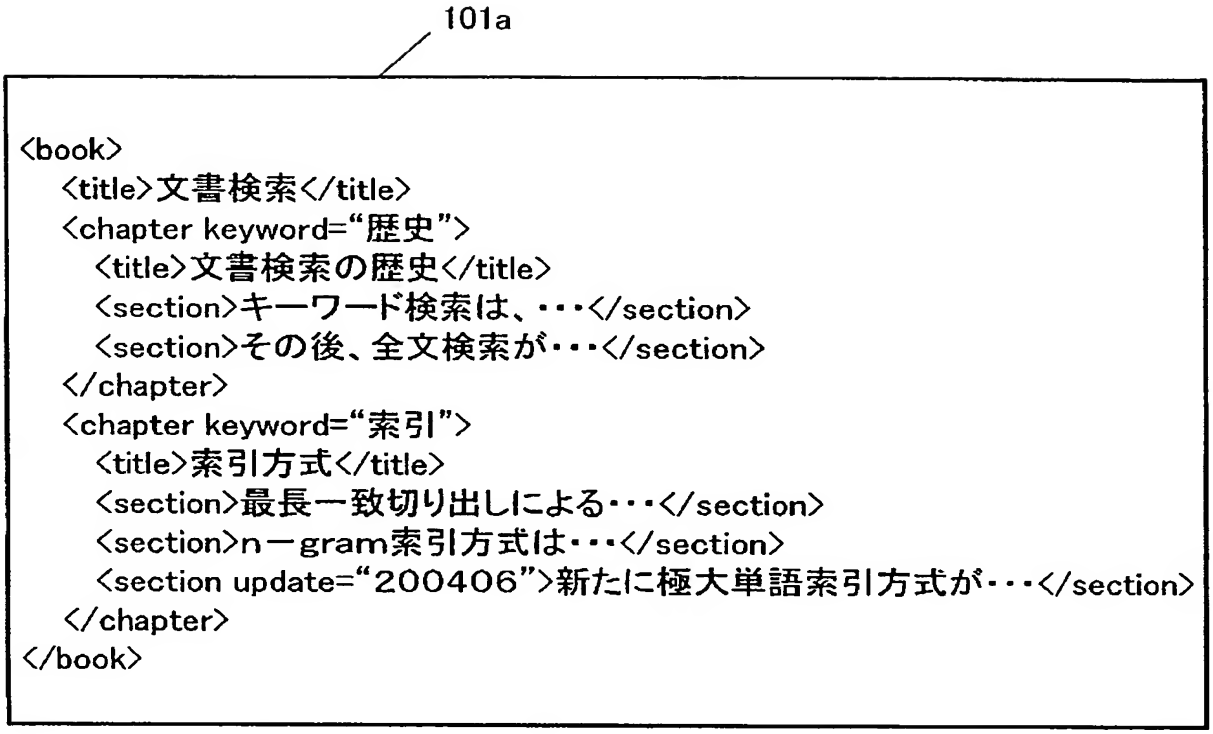
[図1]


構造化
文書群
101
102
入力文書解析部
要素名
登録部
103
祖先パス名
登録部
104
属性名
登録部
105
出現情報登録部
106
要素名
辞書
107
祖先パス名
辞書
108
属性名
辞書
109
要素出現情報格納部
111
祖先パス出現情報
格納部
112
属性出現情報格納部
113
テキスト出現情報
格納部
114
出現位置索引
110
検索条件解析部
117
出現情報取得部
118
検索条件入力部
116
検索結果出力部
119
検索式
115
検索結果
120

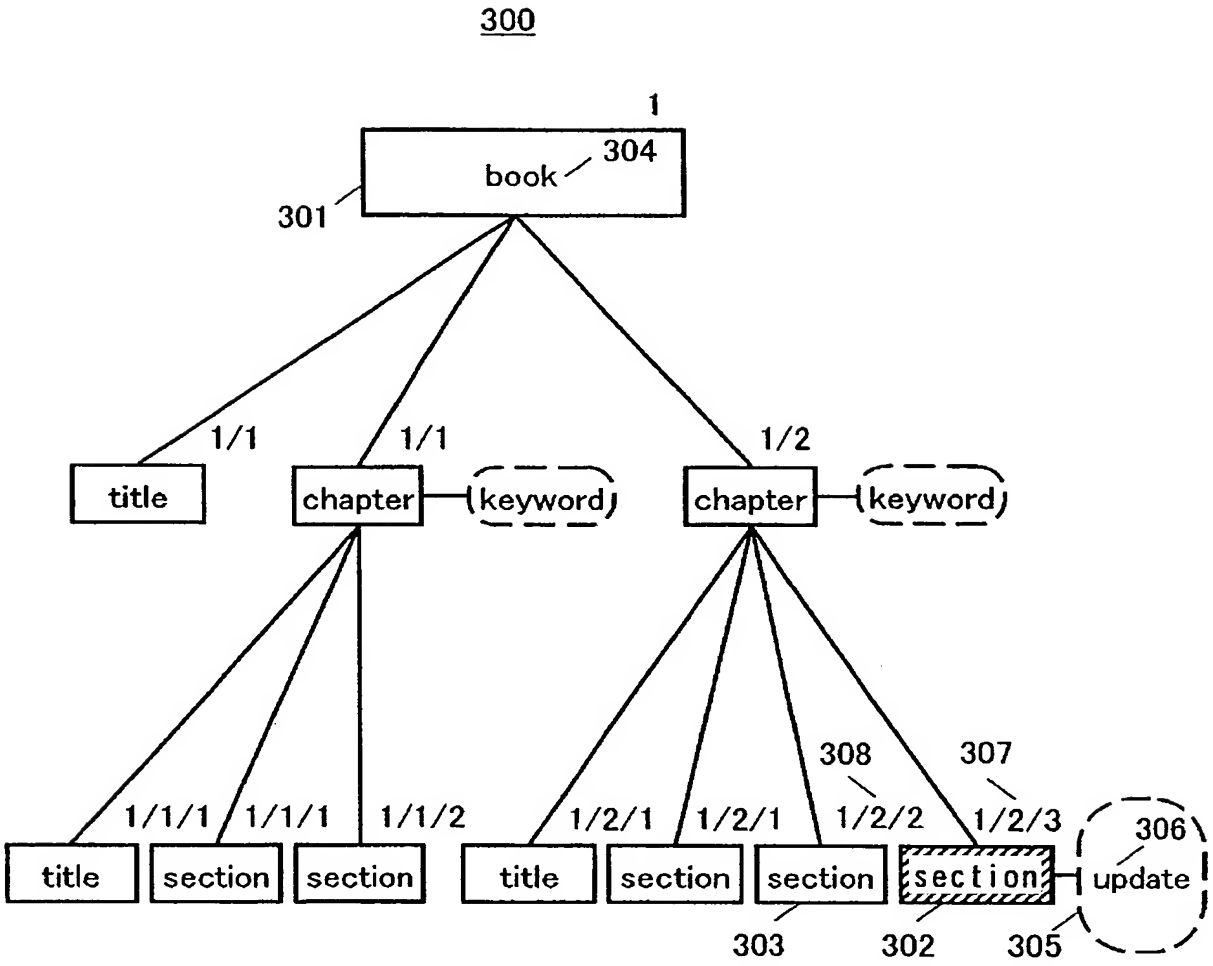
[図2]


開始
入力文書の読込と
文書番号の割り当て
2201
論理構造の解析
2202
要素名の登録
2203
祖先パス名の登録
2204
着目要素に未処理の
属性があるか？
2205
いいえ
はい
属性名の登録
2206
要素出現情報の登録
2207
祖先パス出現情報の登録
2208
着目要素に未処理の
属性があるか？
2209
いいえ
はい
属性出現情報の登録
2210
テキスト出現情報の登録
2211
着目要素に未処理の
属性があるか？
2212
いいえ
はい
テキスト出現情報の登録
（属性値）
2213
全ての要素について
処理したか？
2214
いいえ
はい
全ての入力文書に
ついて処理したか？
2215
いいえ
はい
終了

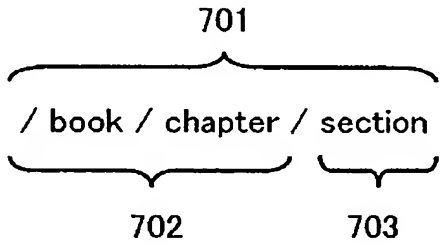
[図3]


101a
<book>
<title>文書検索</title>
<chapter keyword="歴史">
<title>文書検索の歴史</title>
<section>キーワード検索は、・・・</section>
<section>その後、全文検索が・・・</section>
</chapter>
<chapter keyword="索引">
<title>索引方式</title>
<section>最長一致切り出しによる・・・</section>
<section>n－gram索引方式は・・・</section>
<section update="200406">新たに極大単語索引方式が・・・</section>
</chapter>
</book>

[図4]


300
1
book
304
301
1/1
1/1
1/2
title
chapter
keyword
chapter
keyword
1/1/1
1/1/1
1/1/2
1/2/1
1/2/1
1/2/2
1/2/3
308
307
306
title
section
section
title
section
section
section
update
303
302
305

[図5]


701
/ book / chapter / section
702
703

[図6]

407

要素名ID	要素名
1	book
2	title
3	chapter
4	section

[図7]

408

祖先パス名ID	祖先パス名
1	/
2	/book
3	/book/chapter

[図8]

409

属性名ID	属性名
1	keyword
2	update

[図9]

410

411 文字	文	書	検	索	文	書	検	索	の	歴	史
412 文字位置	0	1	2	3	4	5	6	7	8	9	10

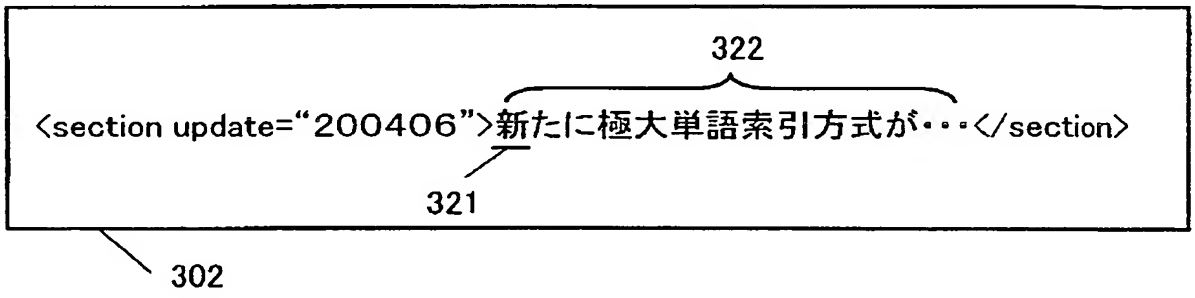
キ	ー	ワ	ー	ド	検	索	は	、	・・・
11	12	13	14	15	16	17	18	19	

[図10A]

501

502 要素名ID	503 文書番号	504 文字位置	505 文字数	506 祖先パス名ID	507 分岐順
4	1	115	40	3	1/2/3

[図10B]



[図11]

511

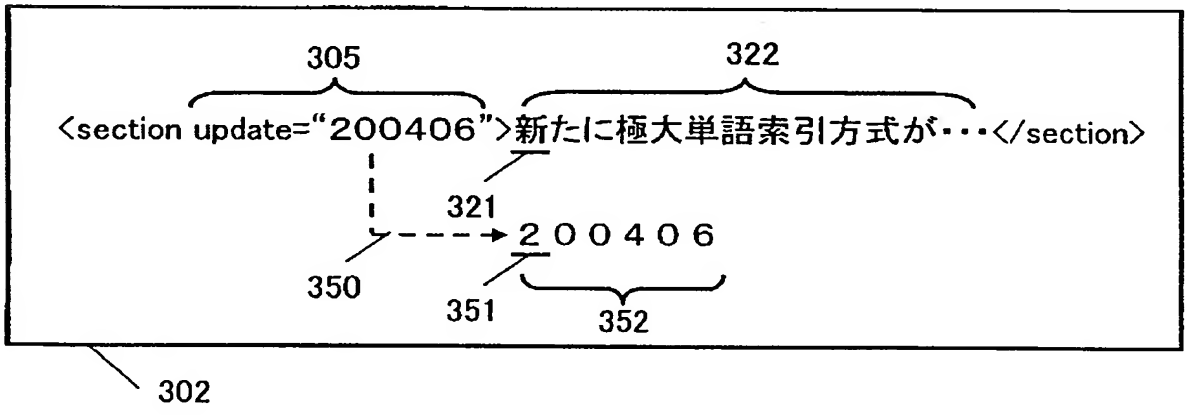
506 祖先パス名ID	503 文書番号	504 文字位置	505 文字数	502 要素名ID	507 分岐順
3	1	115	40	4	1/2/3

[図12A]

521

522 属性名ID	503 文書番号	504 文字位置	505 文字数	506 祖先パス名ID	502 要素名ID	507 分岐順
2	1	115	6	3	4	1/2/3

[図12B]

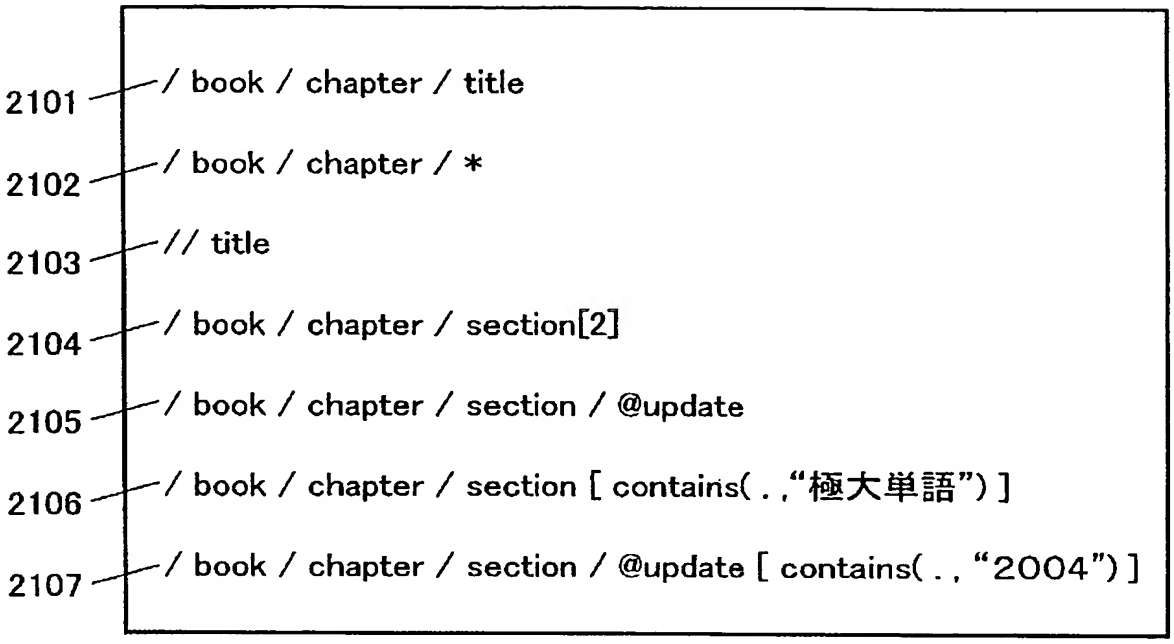


[図13]

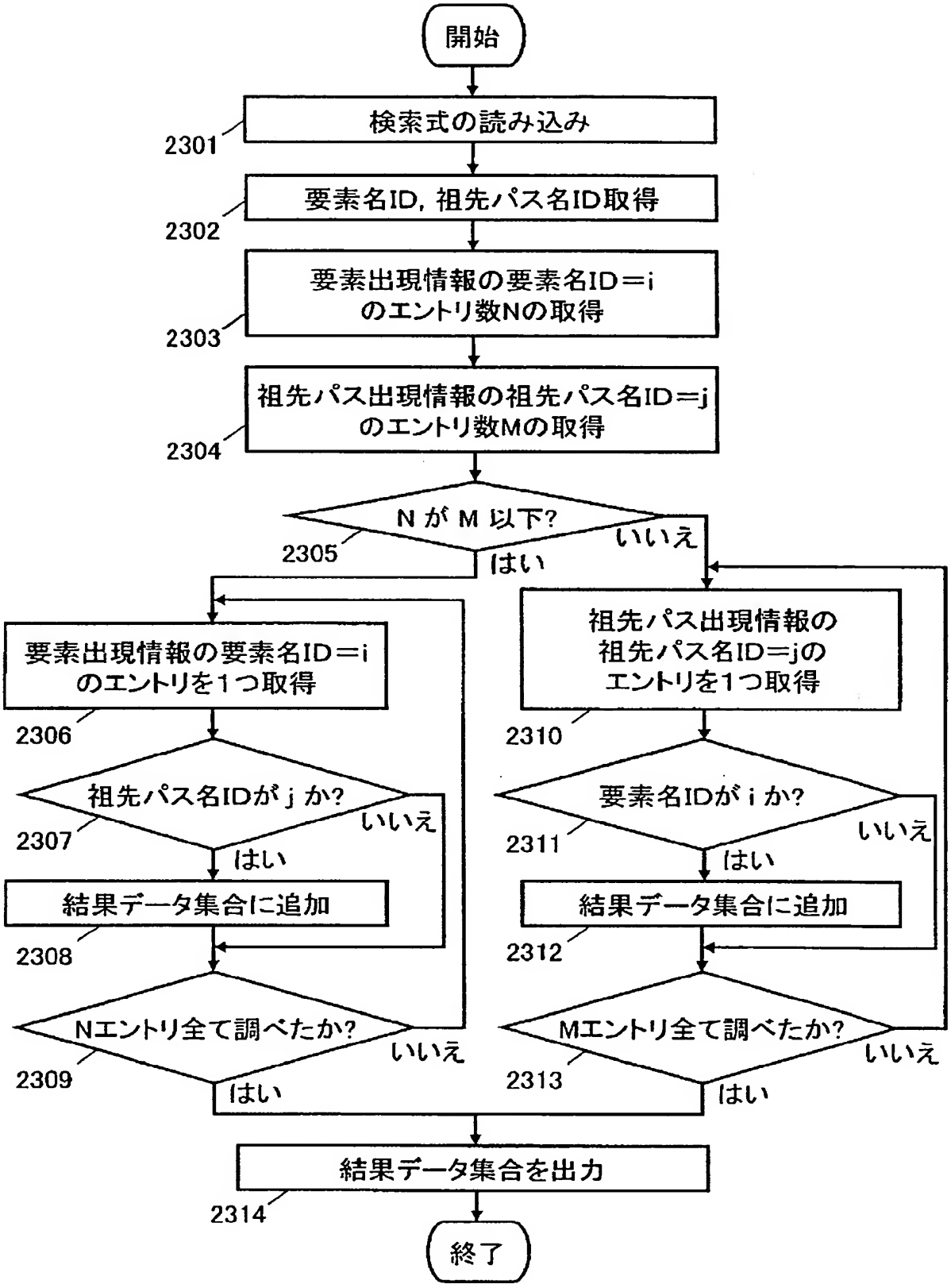
531

532 部分文字列	503 文書番号	504 文字位置	506 祖先パス名ID	502 要素名ID	522 属性名ID	507 分岐順	
“新た”	1	115	3	4	0	1/2/3	
“たに”	1	116	3	4	0	1/2/3	
“に極”	1	117	3	4	0	1/2/3	1201
“極大”	1	118	3	4	0	1/2/3	
“大単”	1	119	3	4	0	1/2/3	
“単語”	1	120	3	4	0	1/2/3	
“20”	1	115	3	4	2	1/2/3	1202
“00”	1	116	3	4	2	1/2/3	
“04”	1	117	3	4	2	1/2/3	
“40”	1	118	3	4	2	1/2/3	
“06”	1	119	3	4	2	1/2/3	

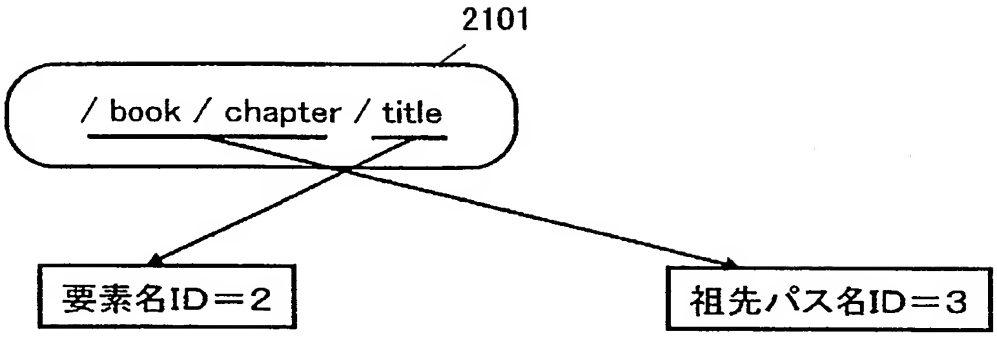
[図14]


2101
/ book / chapter / title
2102
/ book / chapter / *
2103
// title
2104
/ book / chapter / section[2]
2105
/ book / chapter / section / @update
2106
/ book / chapter / section [contains(. ,“極大単語”)]
2107
/ book / chapter / section / @update [contains(. , “2004”)]

[図15]


開始
検索式の読み込み
2301
要素名ID，祖先パス名ID取得
2302
要素出現情報の要素名ID＝i
のエントリ数Nの取得
2303
祖先パス出現情報の祖先パス名ID＝j
のエントリ数Mの取得
2304
N が M 以下?
2305
いいえ
はい
要素出現情報の要素名ID＝i
のエントリを1つ取得
2306
祖先パス名IDが j か?
2307
いいえ
はい
結果データ集合に追加
2308
Nエントリ全て調べたか?
2309
いいえ
はい
祖先パス出現情報の
祖先パス名ID＝jの
エントリを1つ取得
2310
要素名IDが i か?
2311
いいえ
はい
結果データ集合に追加
2312
Mエントリ全て調べたか?
2313
いいえ
はい
結果データ集合を出力
2314
終了

[図16A]



[図16B]

要素名ID	文書番号	文字位置	文字数	祖先パス名ID	分岐順
			・・・		
2	1	0	4	2	1/1
	1	4	7	3	1/1/1
	1	59	4	3	1/2/1
	2	0	6	2	1/1
	4	0	8	2	1/1
	4	24	6	3	1/1/1
	7	0	5	2	1/1
	9	7	4	3	1/1/1
3			・・・		

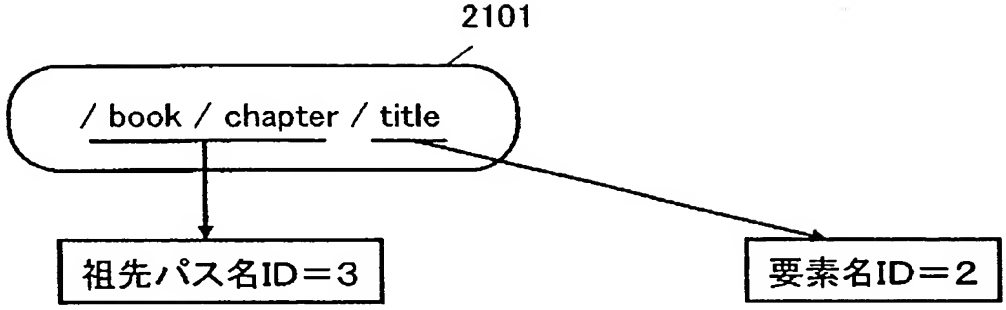
1301

[図16C]

1302

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，2，0，1/1/1），
（1，3，2，0，1/2/1），
（4，3，2，0，1/1/1），
（9，3，2，0，1/1/1）}

[図17A]



[図17B]

祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順
			・・・		
3	1	4	7	2	1/1/1
	1	11	28	4	1/1/1
	1	39	20	4	1/1/2
	1	59	4	2	1/2/1
	1	63	30	4	1/2/1
	1	93	22	4	1/2/2
	1	115	40	4	1/2/3
	3	5	25	4	1/1/1
	4	24	6	2	1/1/1
	4	60	15	4	1/1/1
	6	64	6	4	1/1/1
	9	7	4	2	1/1/1
4			・・・		

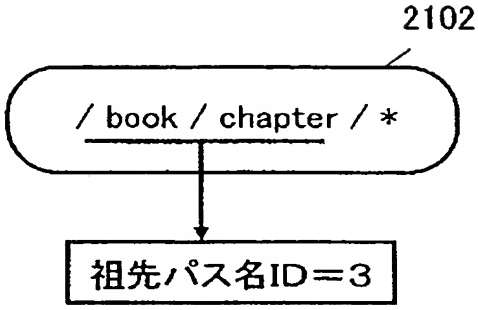
1401

[図17C]

1402

```
(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={(1, 3, 2, 0, 1/1/1),
  (1, 3, 2, 0, 1/2/1),
  (4, 3, 2, 0, 1/1/1),
  (9, 3, 2, 0, 1/1/1)}
```

[図18A]



[図18B]

祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順
			・・・		
3	1	4	7	2	1/1/1
	1	11	28	4	1/1/1
	1	39	20	4	1/1/2
	1	59	4	2	1/2/1
	1	63	30	4	1/2/1
	1	93	22	4	1/2/2
	1	115	40	4	1/2/3
	3	5	25	4	1/1/1
	4	24	6	2	1/1/1
	4	60	15	4	1/1/1
	6	64	6	4	1/1/1
	9	7	4	2	1/1/1
4			・・・		

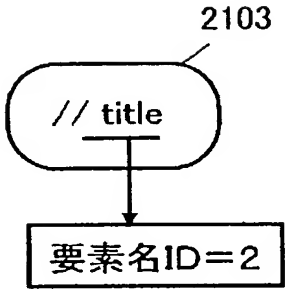
1501

[図18C]

1502

```
(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={(1, 3, 2, 0, 1/1/1),
  (1, 3, 4, 0, 1/1/1),
      ・・・
  (6, 3, 4, 0, 1/1/1),
  (9, 3, 2, 0, 1/1/1)}
```

[図19A]



[図19B]

要素名ID	文書番号	文字位置	文字数	祖先パス名ID	分岐順
			・ ・ ・		
2	1	0	4	2	1/1
	1	4	7	3	1/1/1
	1	59	4	3	1/2/1
	2	0	6	2	1/1
	4	0	8	2	1/1
	4	24	6	3	1/1/1
	7	0	5	2	1/1
	9	7	4	3	1/1/1
3			・ ・ ・		

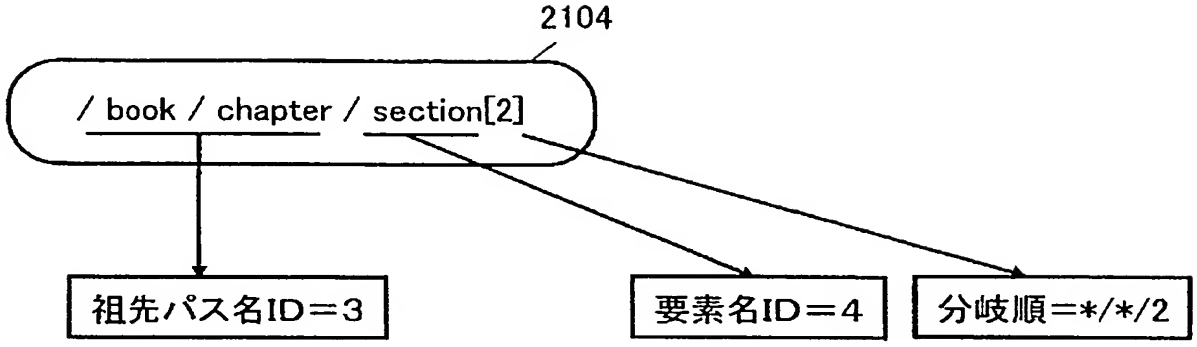
1601

[図19C]

1602

```
(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
=[(1, 2, 2, 0, 1/1  ),
  (1, 3, 2, 0, 1/1/1),
      ・・・
  (7, 2, 2, 0, 1/1  ),
  (9, 3, 2, 0, 1/1/1)]
```

[図20A]



[図20B]

祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順
		1			
3	1	4	7	2	1/1/1
	1	11	28	4	1/1/1
	1	39	20	4	1/1/2
	1	59	4	2	1/2/1
	1	63	30	4	1/2/1
	1	93	22	4	1/2/2
	1	115	40	4	1/2/3
	3	5	25	4	1/1/1
	4	24	6	2	1/1/1
	4	60	15	4	1/1/1
	6	64	6	4	1/1/1
	9	7	4	2	1/1/1
4					

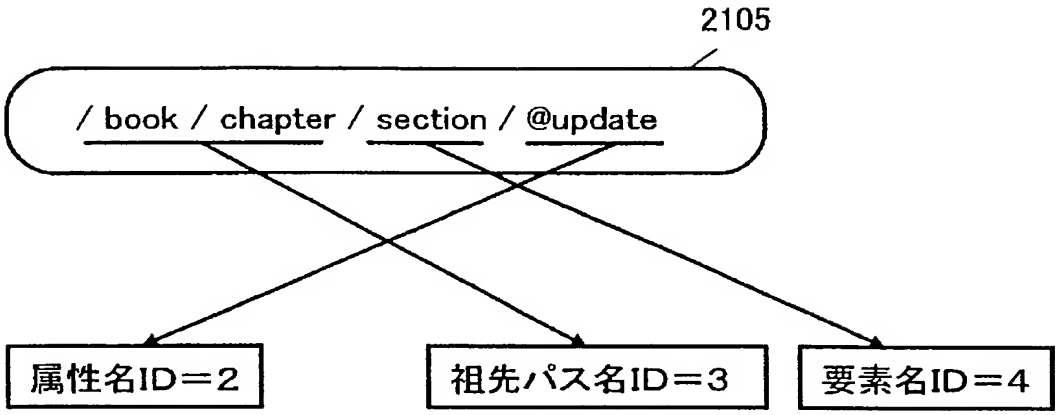
1701

[図20C]

1702

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，4，0，1/1/2），
（1，3，4，0，1/2/2）}

[図21A]



[図21B]

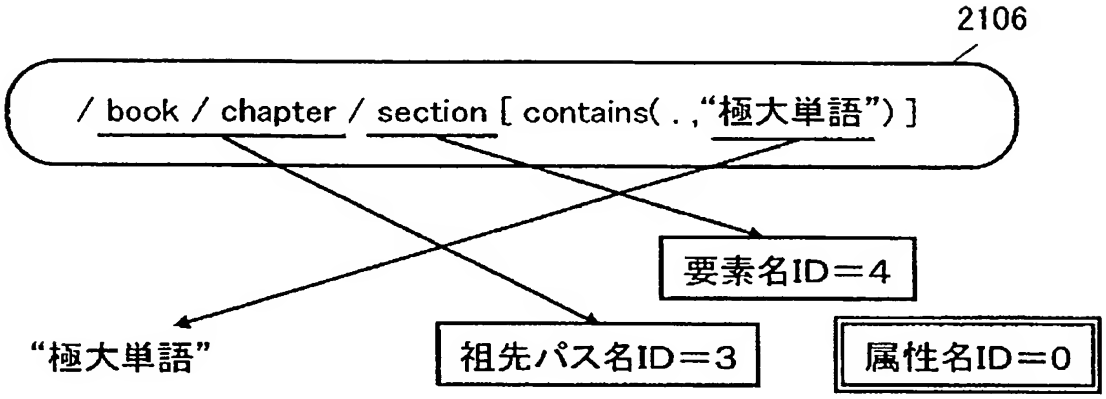
属性名ID	文書番号	文字位置	文字数	祖先パス名ID	要素名ID	分岐順	
			・ ・ ・				
2	1	115	6	3	4	1/2/3	1801
	2	8	4	2	2	1/1	
	5	60	6	3	4	1/1/2	
	8	32	8	3	2	1/2/1	
3			・ ・ ・				

[図21C]

1802

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
=[(1, 3, 4, 2, 1/2/3),
(5, 3, 4, 2, 1/1/2)]

[図22A]



[図22B]

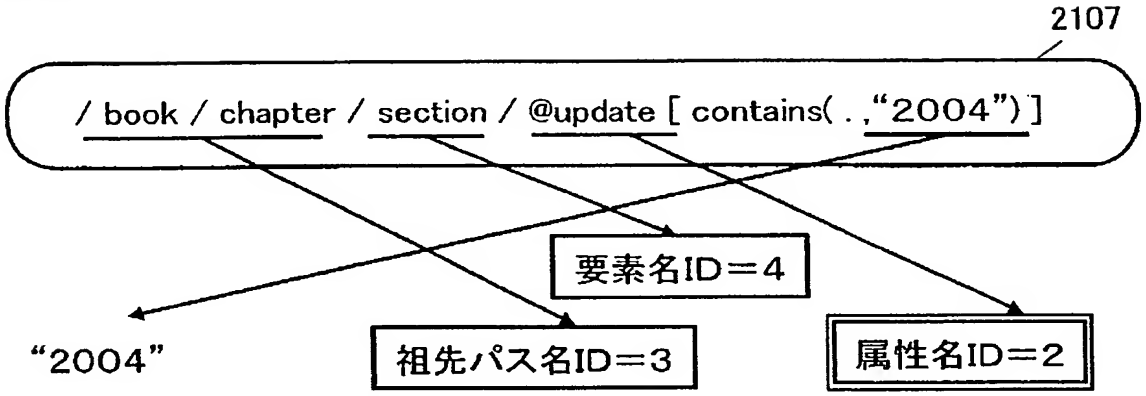
部分文字列	文書番号	文字位置	祖先パス名ID	要素名ID	属性名ID	分岐順	
			・				
			・				
			・				
“極大”	1	118	3	4	0	1/2/3	1901
	2	86	3	4	0	1/1/1	
	3	24	2	2	0	1/1	
	4	62	3	4	0	1/1/1	
	8	77	3	4	2	1/1/1	
			・				
			・				
			・				
“単語”	1	120	3	4	0	1/2/3	1902
	1904 3	26	2	2	0	1/1	
	4	64	3	4	0	1/1/1	
	1905 8	79	3	4	1	1/1/1	
			・				
			・				
			・				

[図22C]

1903

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，4，0，1/2/3），
（4，3，4，0，1/1/1）}

[図23A]



[図23B]

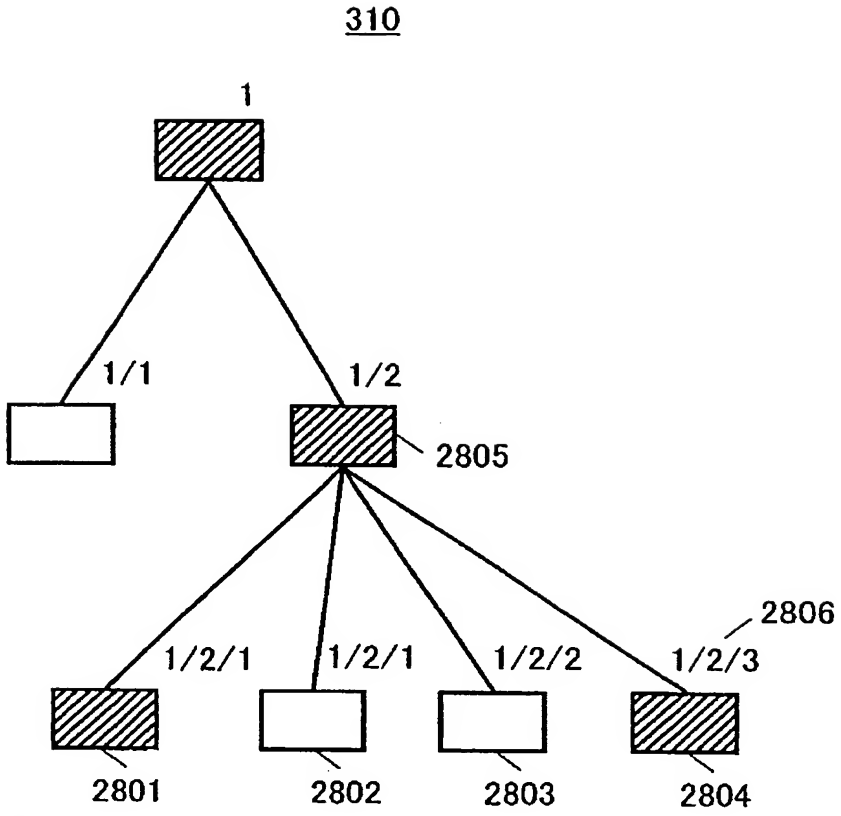
部分文字列	文書番号	文字位置	祖先パス名ID	要素名ID	属性名ID	分岐順	
“20”	1	115	3	4	2	1/2/3	2001
	2	15	3	4	0	1/1/1	
	3	24	2	2	0	1/1	
	5	21	3	4	2	1/1/1	
	7	54	3	4	1	1/1/1	
“04”	1	117	3	4	2	1/2/3	2002
	2004 3	26	2	2	0	1/1	
	5	23	3	4	2	1/1/1	
	2005 7	56	3	4	1	1/1/1	

[図23C]

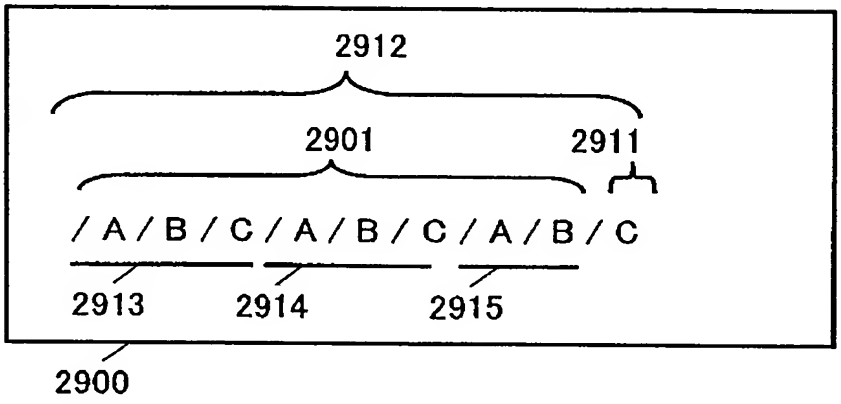
2003

(文書番号, 祖先パス名ID, 要素名ID, 属性名ID, 分岐順)
={(1, 3, 4, 2, 1/2/3),
(5, 3, 4, 2, 1/1/1)}

[図24]


310
1
1/1
1/2
2805
2806
1/2/1
1/2/1
1/2/2
1/2/3
2801
2802
2803
2804

[図25A]


2912
2901
2911
/ A / B / C / A / B / C / A / B / C
2913
2914
2915
2900

[図25B]

2903

2904 祖先パスID	2905 祖先パス名
1	/
・・・	・・・
25	/A/B
・・・	・・・
83	/A/B/C
・・・	・・・

[図25C]

祖先パス名ID列＝83:83:25

2902

[図26]

541

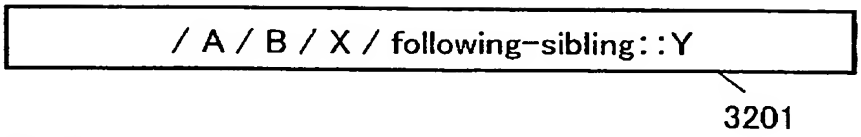
502 要素名ID	503 文書番号	504 文字位置	505 文字数	506 祖先パス名ID	507 分岐順	548 空要素順
10	1	100	20	83:25	1/2/3/1/1/2	1/1/2/1/2/1

[図27]

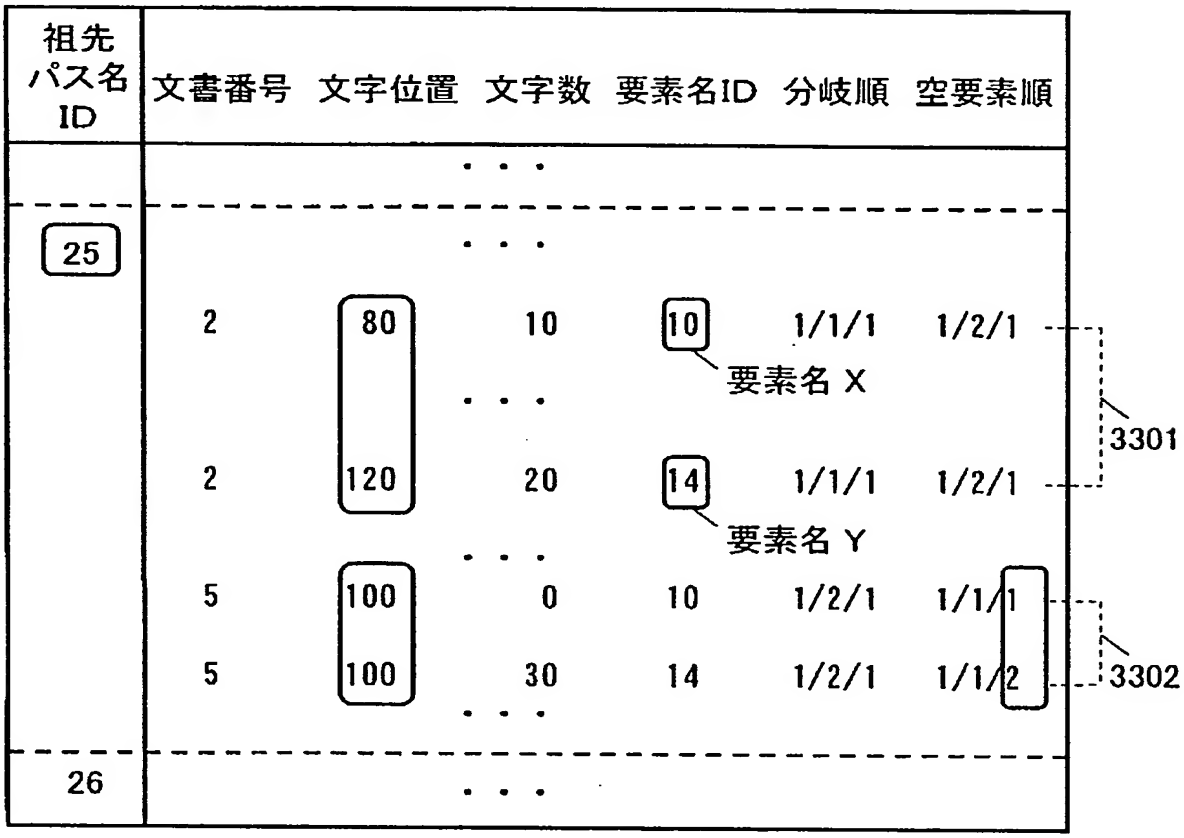
551

506 祖先パス名ID	503 文書番号	504 文字位置	505 文字数	502 要素名ID	507 分岐順	548 空要素順
83:25	1	100	20	10	1/2/3/1/1/2	1/1/2/1/2/1

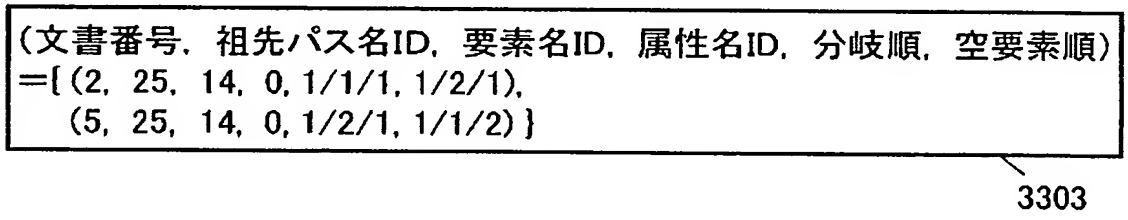
[図28]



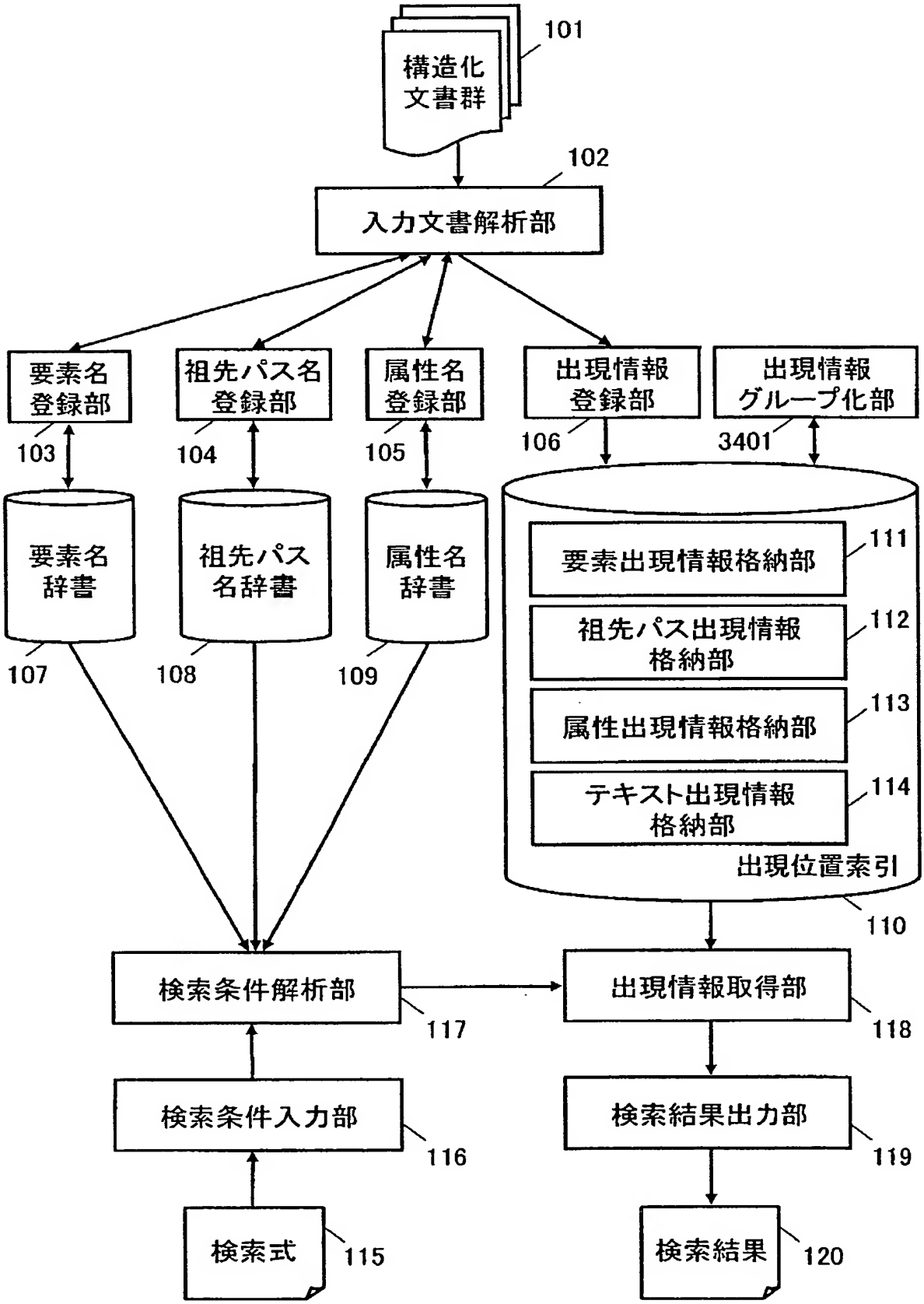
[図29A]



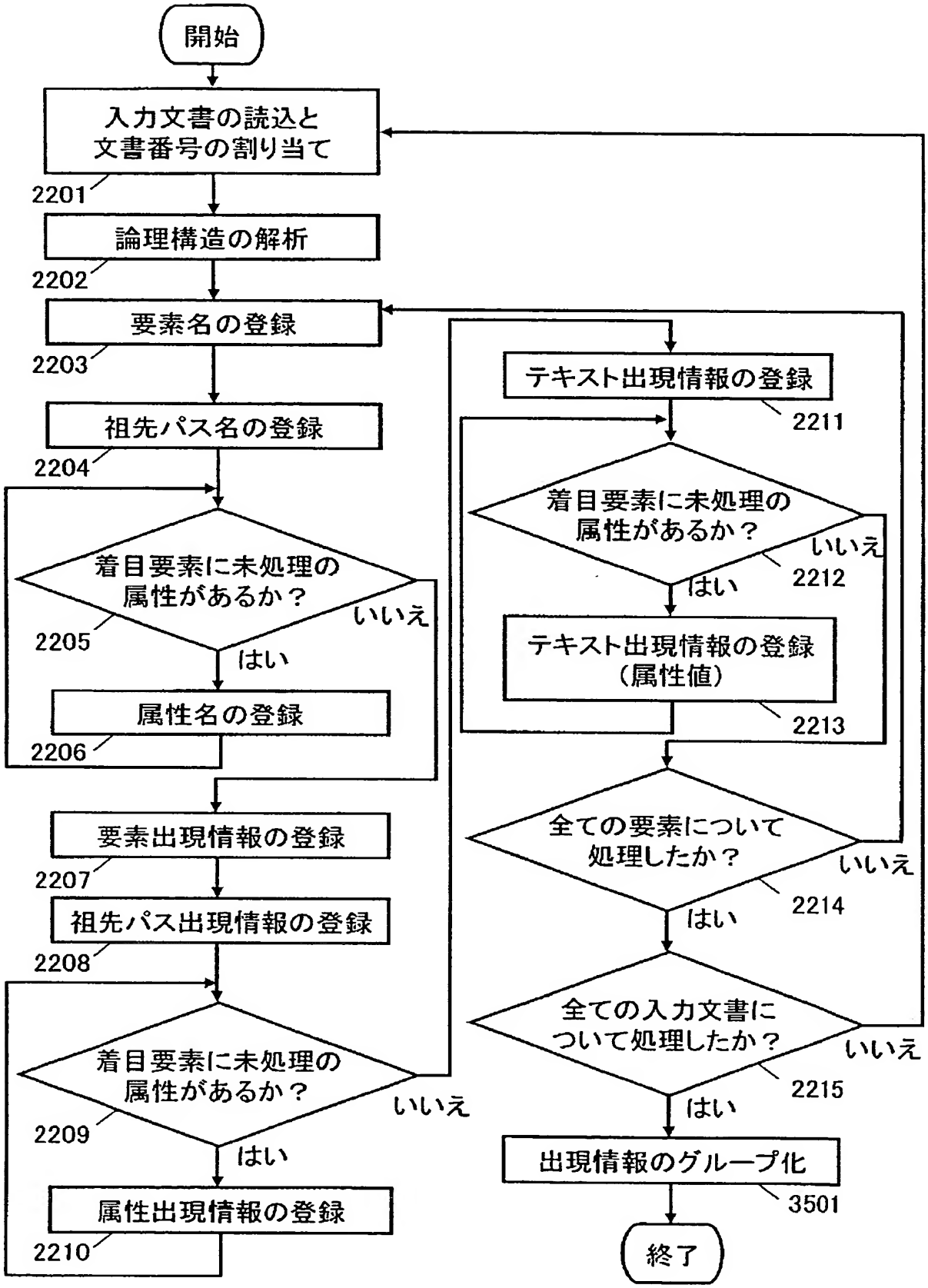
[図29B]



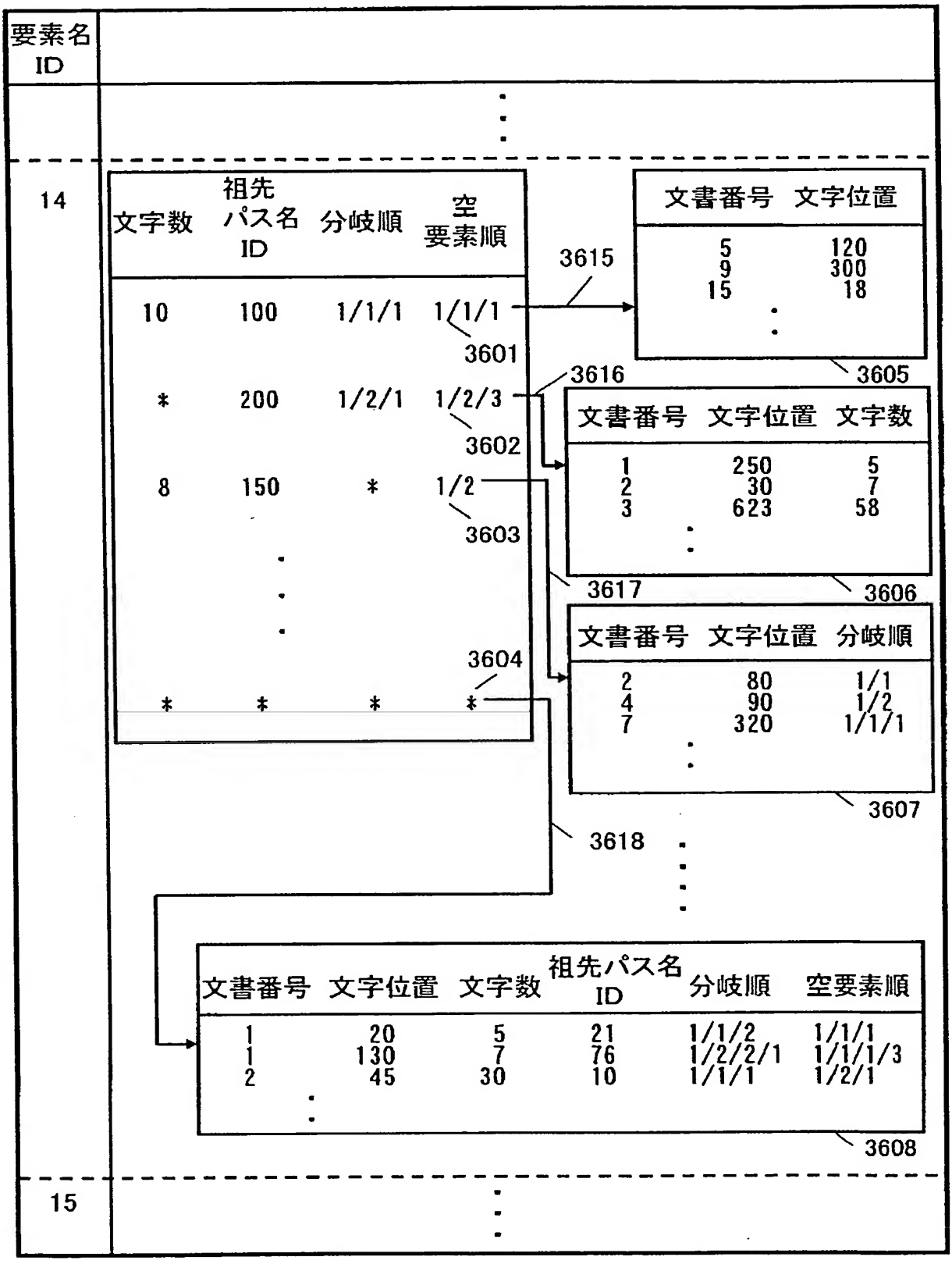
[図30]


構造化文書群
101
102
入力文書解析部
要素名登録部
103
祖先パス名登録部
104
属性名登録部
105
出現情報登録部
106
出現情報グループ化部
3401
要素名辞書
107
祖先パス名辞書
108
属性名辞書
109
要素出現情報格納部
111
祖先パス出現情報格納部
112
属性出現情報格納部
113
テキスト出現情報格納部
114
出現位置索引
110
検索条件解析部
117
出現情報取得部
118
検索条件入力部
116
検索結果出力部
119
検索式
115
検索結果
120

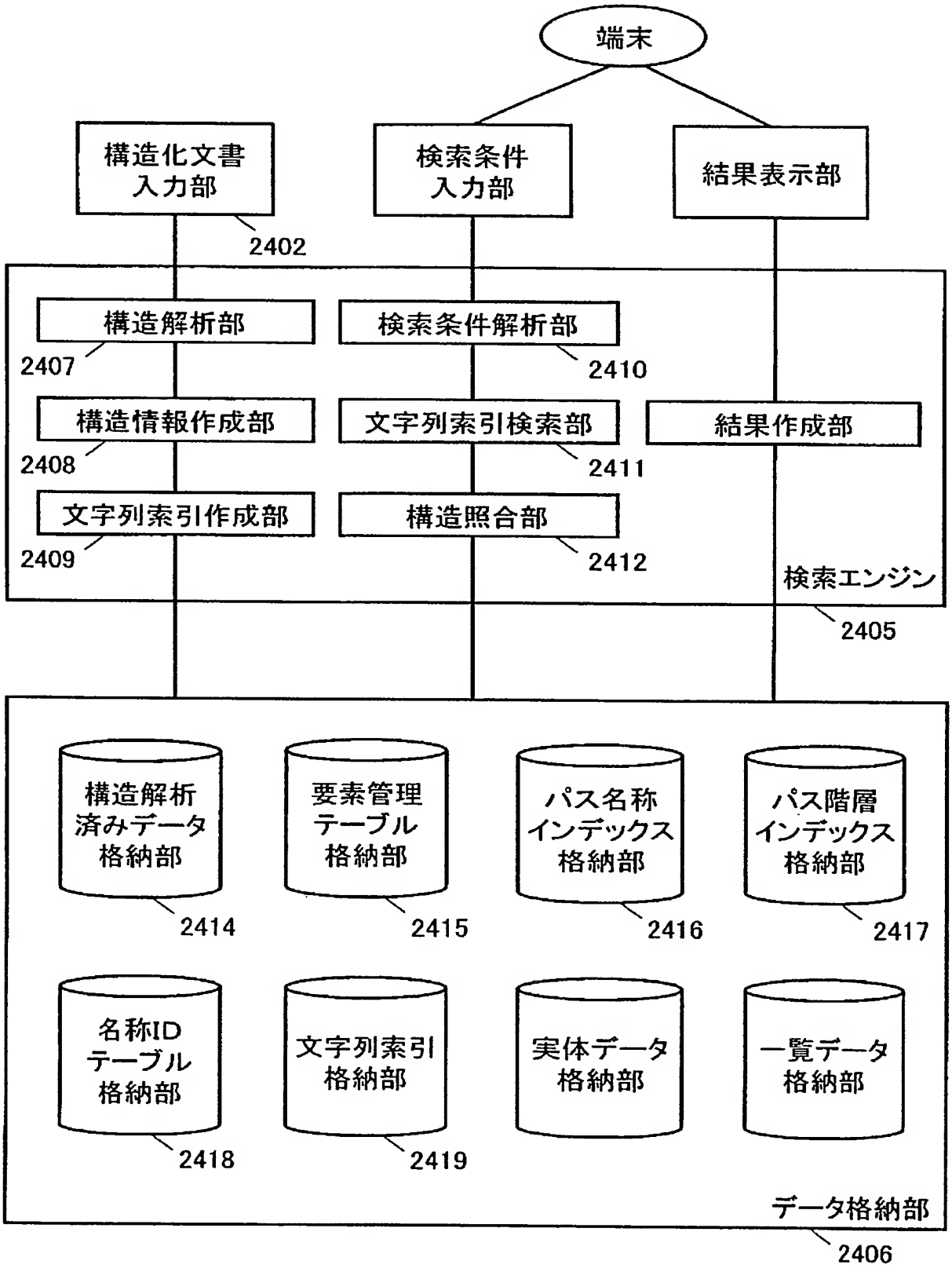
[図31]


開始
入力文書の読込と
文書番号の割り当て
2201
論理構造の解析
2202
要素名の登録
2203
祖先パス名の登録
2204
着目要素に未処理の
属性があるか？
2205
いいえ
はい
属性名の登録
2206
要素出現情報の登録
2207
祖先パス出現情報の登録
2208
着目要素に未処理の
属性があるか？
2209
いいえ
はい
属性出現情報の登録
2210
テキスト出現情報の登録
2211
着目要素に未処理の
属性があるか？
2212
いいえ
はい
テキスト出現情報の登録
（属性値）
2213
全ての要素について
処理したか？
2214
いいえ
はい
全ての入力文書に
ついて処理したか？
2215
いいえ
はい
出現情報のグループ化
3501
終了

[図32]


要素名ID
14
15
文字数
祖先パス名ID
分岐順
空要素順
10 100 1/1/1 1/1/1
* 200 1/2/1 1/2/3
8 150 * 1/2
* * * *
3601
3602
3603
3604
3615
3616
3617
3618
文書番号 文字位置
5 120
9 300
15 18
3605
文書番号 文字位置 文字数
1 250 5
2 30 7
3 623 58
3606
文書番号 文字位置 分岐順
2 80 1/1
4 90 1/2
7 320 1/1/1
3607
文書番号 文字位置 文字数 祖先パス名ID 分岐順 空要素順
1 20 5 21 1/1/2 1/1/1
1 130 7 76 1/2/2/1 1/1/1/3
2 45 30 10 1/1/1 1/2/1
3608

[図33]


端末
構造化文書
入力部
検索条件
入力部
結果表示部
2402
構造解析部
検索条件解析部
2407
2410
構造情報作成部
文字列索引検索部
結果作成部
2408
2411
文字列索引作成部
構造照合部
2409
2412
検索エンジン
2405
構造解析
済みデータ
格納部
要素管理
テーブル
格納部
パス名称
インデックス
格納部
パス階層
インデックス
格納部
2414
2415
2416
2417
名称ID
テーブル
格納部
文字列索引
格納部
実体データ
格納部
一覧データ
格納部
2418
2419
データ格納部
2406

[図34]

2501

2502 検索単位識別子	2503 文書番号	2504 パス名称ID	2505 パス階層ID	2506 名称ID
1	1	N2	L2	T3
2	1	N3	L2	T4
3	1	N3	L6	T4
4	1	N4	L2	T5
5	1	N7	L3	T8
6	1	N8	L3	T9
7	1	N10	L4	T11
8	1	N11	L4	T9
・ ・ ・	・・・	・・・	・・・	・・・

[図35A]

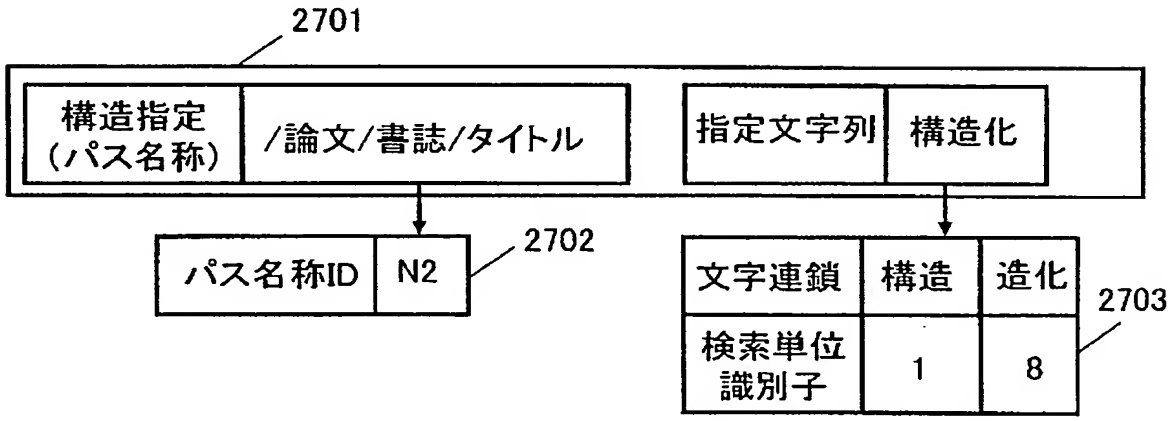
<タイトル>構造化文書管理</タイトル>

[図35B]

2602

文字連鎖 (2603)	検索単位識別子 (2604)	文字位置番号 (2605)
“構造” (2606)	1	1
“造化”	1	2
“化文”	1	3
“文書”	1	4
“書管”	1	5
“管理”	1	6

[図36A]



[図36B]

2501

検索単位識別子	文書番号	パス名称ID	パス階層ID	名称ID
(1)	1	(N2)	L2	T3
:	...	...	...	...
(8)	1	N11	L4	T9
:	...	...	...	...

[図36C]

検索単位識別子	文書番号
1	1